Multivariate time series modelling

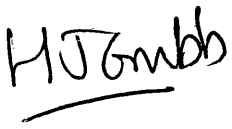
Grubb, Howard John

Award date:
1990

Awarding institution:
University of Bath

Multivariate Time Series Modelling

Submitted by Howard John Grubb
for the degree of PhD
of the University of Bath
1990



H.J.Grubb

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Multivariate Time Series Modelling

Abstract

Multivariate time series data consist of observations on several variables collected sequentially through time. They occur in many fields of study including economic, environmental and industrial applications. They are usually studied in order to explore dynamic relationships between the variables and to summarise these with a small number of parameters in a suitable stochastic model. This model may help to describe and perhaps explain the behaviour of the system under study and, if the system is assumed to remain in its current state, or evolve in a describable way, the model can be used to forecast the variables of the system into the future.

This thesis concentrates on the specification of suitable models for multivariate time series data. We will introduce the general VARMA model, which can be used to describe a variety of dynamic relationships, and outline some of the problems often encountered when trying to build such models for data. These include the presence of co–integration, or common trends in the data, the difficulty of identifying or choosing a suitable model from the general class and the large number of parameters which then need to be estimated. Published approaches aimed at tackling some of these problems, together with some extensions to them will be presented and discussed in detail. Case studies using some published datasets serve to illustrate and assess the effectiveness of the methods and based upon these results we can make some recommendations on the use of such methods. We present a strategy which can be used when building VARMA models for multivariate time series data.

Acknowledgements

This work was carried out under the supervision of Dr Chris Chatfield and I am extremely grateful to him for his help, advice and encouragement throughout.

The research was funded by the Science and Engineering Research Council and I am grateful for their financial support.

I would also like to thank the School of Mathematical Sciences at the University of Bath for their excellent facilities and assistance which made this work possible.

Table of Contents

Notation

§	Symbol	Meaning
	$\underset{k\times m}{A}$	$=\{A_{ij}\}=\{[A]_{ij}\}_{j=1,\ldots,m}^{i=1,\ldots,k}$, $k\times m$ matrix A
	A^T	$=\{A_{ji}\}_{j=1,\ldots,m}^{i=1,\ldots,k}$, transpose of matrix A
	$\lvert A\rvert$	determinant of (square) matrix A
	A^{-1}	inverse of (square) matrix A
	tr$A$	$=\sum_{i=1}^{k}A_{ii}$ = trace of (square) matrix $\underset{k\times k}{A}$
	vec$A$	$=(A_{\cdot 1}^T,\ldots,A_{\cdot m}^T)^T$, vector made out of the columns of matrix $\underset{k\times m}{A}$
	$A\otimes B$	Kronecker product of matrices A and B, $=\{A_{q_kq_m}B_{r_nr_p}\}_{j=1,\ldots,pm}^{i=1,\ldots,kn}$; $\underset{k\times m}{A}$, $\underset{n\times p}{B}$, $i=q_kk+r_k=q_nn+r_n$, $j=q_mm+r_m=q_pp+r_p$.
	I_k	$k\times k$ identity matrix
	$\underset{k\times m}{0}$	$k\times m$ matrix of zeroes
	diag(a_i)	diagonal matrix A, $A_{ii}=a_i$
	$[\underset{k\times m}{A} : \underset{k\times n}{B}]$	matrix C, $C_{ij}=\begin{cases}A_{ij} & j\le m\\ B_{i,j-m} & m<j\le n\end{cases}$
	ln	natural logarithm
	argmin	argmin$f(x)=x_0 \Rightarrow f(x)>f(x_0)\ \forall x\neq x_0$
	δ_{ij}	$=\begin{cases}1 \text{ if } i=j\\ 0 \text{ if } i\neq j\end{cases}$ Kronecker delta
	$N(\mu,\sigma^2)$	normal (Gaussian) random variable, mean μ, variance σ^2
	$N_k(\mu,\Sigma)$	k–variate normal, mean μ, covariance matrix Σ
	χ_ν^2	χ^2 distribution with ν degrees of freedom
	$\chi_{\alpha,\nu}^2$	upper α% point of a χ_ν^2 distribution
	var/cov	variance/covariance
	$A\backslash B$	$=\{x:x\in A,x\notin B\}$, set difference operation
1.1.2	VARMA	vector autoregressive moving–average (model)
1.2	ARMA	autoregressive moving–average (model)
1.2.1	z_t	observation on time series variable at time t
1.2.1	a_t	white noise series at time t (i.i.d. $\sim N(0,\sigma_a^2)$)
1.2.1	B	backshift operator: $Bz_t = z_{t-1}$
1.2.1	$\varphi(B)$	$=1-\varphi_1B-\ldots-\varphi_pB^p$, univariate autoregressive operator
1.2.1	$\theta(B)$	$=1-\theta_1B-\ldots-\theta_qB^q$, univariate moving average operator
1.2.2	ρ_h	univariate correlation between z_t and z_{t+h} (acf)
1.2.2	φ_{hh}	univariate partial autocorrelation (pacf)
1.3.1	TFN	transfer function noise (model)

1.3.1	x_t, y_t, n_t	input, output, noise series
1.3.1	$\nu(B)$	$=\nu_0+\nu_1 B+\nu_2 B^2+\ldots$, transfer function (operator)
		ν_i= impulse response weights
1.3.1	$\left.\begin{matrix}\omega(B)\\ \delta(B)\end{matrix}\right\}$	polynomials used to represent $\nu(B)$ in rational form
1.3.2	$\rho_{xy}(h)$	correlation between x_t and y_{t+h} (ccf)
2.1	$\mathbf{z}_t$	$=(z_{1t},\ldots,z_{kt})^T$, $k\times 1$ vector of observations at time t
2.1	$\mathbf{a}_t$	$k\times 1$ vector of white noise (i.i.d. $\sim N_k(\mathbf{0},\Sigma_a)$)
2.1	$\Phi(B)$	$=I-\Phi_1 B-\ldots-\Phi_p B^p$, $k\times k$ AR matrix polynomial
2.1	$\Theta(B)$	$=I-\Theta_1 B-\ldots-\Theta_q B^q$, $k\times k$ MA matrix polynomial
2.1.1	$\Psi(B)$	$=\Psi_0+\Psi_1 B+\ldots$, MA(∞) representation of process
2.1.2	$\mathbf{x}_t$	vector of exogenous variables
2.1.3	$\Pi(B)$	$=I-\Pi_1 B-\ldots$, AR(∞) representation of process
2.2.1	$\Gamma(h)$	$=\mathrm{cov}(\mathbf{z}_t,\mathbf{z}_{t+h})$, $k\times k$ covariance matrix of $\mathbf{z}_t$, $\mathbf{z}_{t+h}$ (ccv)
	Γ_h	ccv in §5.2.4
2.2.1	$\rho(h)$	correlation between $\mathbf{z}_t$, $\mathbf{z}_{t+h}$ (ccf)
2.2.2	$\mathcal{P}(h)$	partial autoregression matrix at lag h (parm)
2.2.2	$M(p)$	sample statistic for parm – see [2.18]
2.2.2	$Q(h)$	partial autocorrelation function (pccf)
2.2.2	$P(h)$	partial lag autocorrelation function (plcm)
2.2.2	$X(h)$	sample statistic for plcm – see [2.23]
2.2.3	$X \xrightarrow{P} Y$	$P(\lvert X(n)-Y\rvert>\varepsilon)\to 0$ as $n\to\infty$ $\forall\varepsilon>0$
2.5	$\hat{\mathbf{z}}_t(h)$	forecast of $\mathbf{z}_{t+h}$ made at time t
2.5	$\hat{\mathbf{e}}_t(h)$	$=\mathbf{z}_{t+h}-\hat{\mathbf{z}}_t(h)$, error in forecast of $\mathbf{z}_{t+h}$ made at t
2.5	$V(h)$	covariance matrix of h–step ahead forecast errors
3.1.1	$g(z_t,\lambda)$	$=z_t^{(\lambda)}=\dfrac{z_t^{\lambda}-1}{\lambda}$, power transformation
3.1.1	∇	differencing operator: $\nabla=1-B$, $\nabla\mathbf{z}_t=\mathbf{z}_t-\mathbf{z}_{t-1}$
3.1.1	I(d)	integrated to degree d
3.1.2	$D(B)$	multivariate differencing operator
3.1.3	$S_t^{(T)}$	step intervention variable at time T
3.1.3	$P_t^{(T)}$	pulse intervention variable at time T
3.3	d(X,Y)	Piccolo's model metric
3.3.1	AIC	Akaike's Information Criterion
3.3.2	MSE	Mean Square Error
3.3.2	GMSE	Geometric Mean Square Error
4.1	n.s.	non–stationary
4.1	CI	Co–Integration/ed/ing (also CI(1,1))
4.1	CI(d,b)	z_{it} are I(d) and $\exists\boldsymbol{\alpha}$ s.t. $\boldsymbol{\alpha}^T\mathbf{z}_t$ is I($d-b$), $b\le d$

4.1.1	E&G	Engle and Granger (1987)
4.4.1	$\hat{DW}$	Durbin–Watson test statistic
4.4.1	$\hat{DF}$	Dickey–Fuller test statistic
4.4.1	$\hat{ADF}(p)$	Augmented Dickey–Fuller test statistic (order p)
5.2	T&T	Tiao and Tsay (1989)
5.2.1	SC	Scalar Component
5.2.1	SCM	Scalar Component Model
5.2.4	Γ_h	ccv
5.2.4	$\Gamma(m,h,j)$	constructed covariance matrix of T&T
5.2.5	$R(m,h,j)$	root table
5.2.5	$D(m,h,j)$	diagonal increment of $R(m,h,j)$
5.2.8	$\mathbf{Y}_{m,t}$	$=(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t-m}^T)^T$, constructed variable of T&T
5.2.8	$\left.\begin{matrix} C(s) \\ crit(s) \end{matrix}\right\}$	statistics of T&T
5.3	T89a	Tsay (1989a)
5.3.1	ψ_t	σ–field generated by $\{\mathbf{a}_t,\mathbf{a}_{t-1},\ldots\}$
5.3.1	KI	Kronecker Index
5.3.1	$\mathbf{z}_{t\mid t-1}$	$=E[\mathbf{z}_t \mid \psi_{t-1}]$, forecast of $\mathbf{z}_t$ made at time $t-1$
5.3.1	H_h	constructed covariance matrix
5.3.1	$\mathbf{F}_{m,t}$	$=(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t+m}^T)^T$, constructed future vector
5.3.2	$\mathbf{F}_t^*$	subvector of $\mathbf{F}_{m,t}$
5.3.4	EFC	Eventual Forecasting Component
5.4	T89b	Tsay (1989b)

Chapter 1. Introduction

1.1 Time series data

Time series data consist of observations on a variable collected sequentially through time. In this thesis we consider the relationships between several series.

It is usually assumed (and mostly the case) that the time intervals between the observations are equal, which simplifies the handling of the data, although we can consider irregularly–spaced observations. Examples of time series data are an industrial process parameter recorded every hour, the daily mean temperature at a particular site, monthly interest rates in an economy and the annual sales of a product – Figure 1.1a displays the 54 observations of the annual dollar sales of the Lydia Pinkham vegetable compound (see Appendix A.1 for a description of this data) plotted against the years in which they were recorded – 1907 to 1960. Such data are usually studied with a view to (a) describing features of the past behaviour of the variable (extrema, growth rates, variability etc. – see e.g. Mills (1990, Part 1)), (b) formulating mathematical models which capture features of interest (for example a linear trend which describes the long term changes in the variable) and (c) extrapolating these models to produce an estimate (or forecast) of the unobserved future values of the variable – Fig 1.1a also shows some possible forecasts of the future sales after 1960. Chatfield (1989), Granger and Newbold (1986), Harvey (1989) and Wei (1990) provide discussions of the nature and aims of time series analysis.

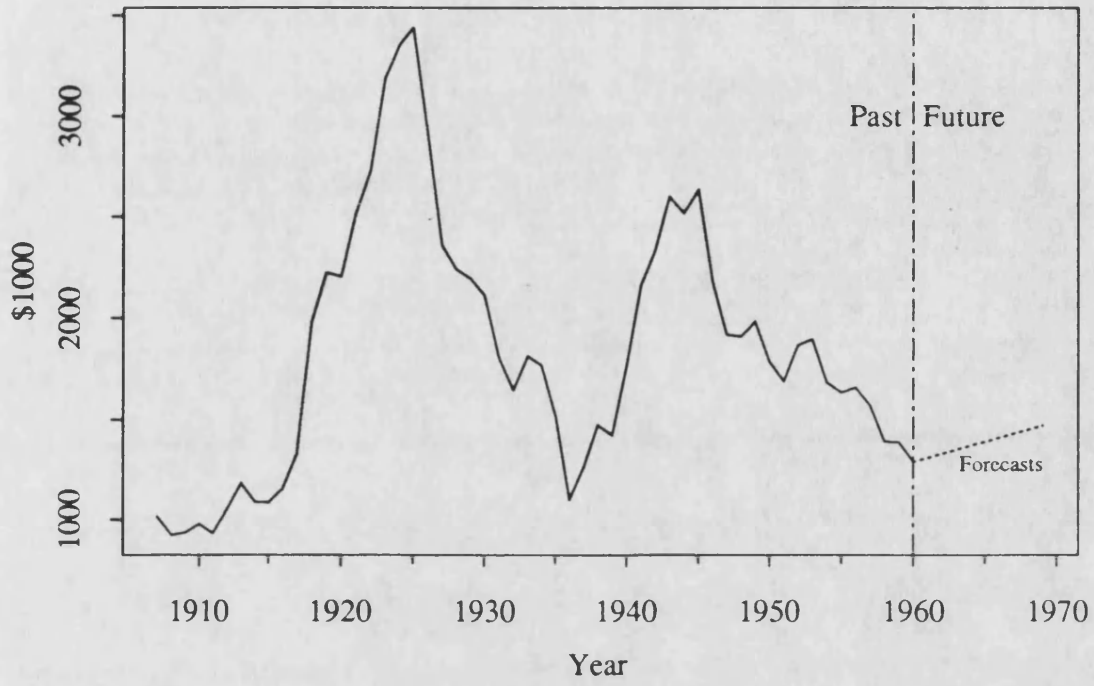
There may be occasions when other objectives are important and perhaps the ordering of the data is not of interest – for instance, a variable such as the temperature of an industrial process may be measured through time, but only to obtain an estimate of its mean level and variability – however in this thesis we are only concerned with the time series nature of variables, in which case the ordering is of prime importance. Implicit in the time–ordering of the variable is the assumption that the future cannot affect the past – this differs from for instance, the spatial context (e.g. the same variable measured at different sites) where there is usually no such notion of causation (except perhaps when the sites have a natural ordering such as different measuring stations along a flowing river – although this could also be thought of in the time series context).

1.1.1 Multivariate time series

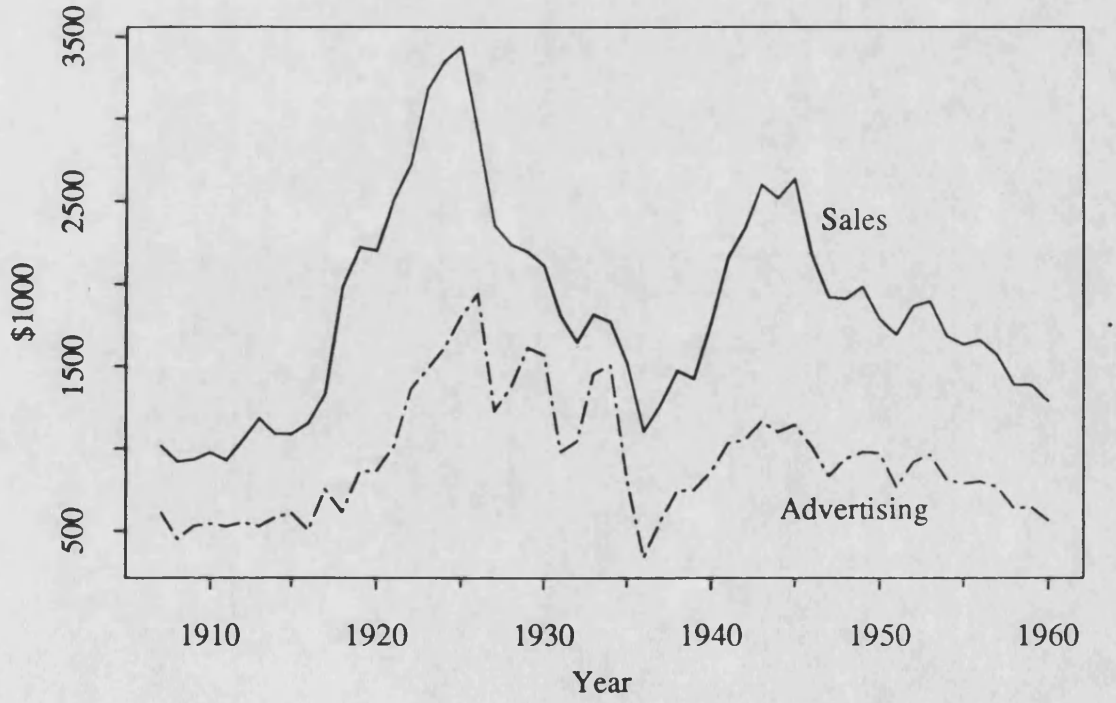
Multivariate time series data consist of observations made simultaneously on several related variables of interest. For example, we may monitor several industrial process parameters every hour, related environmental variables every

Figure 1.1

(a) Lydia Pinkham sales data



(b) Lydia Pinkham data



day or, in the case of the Lydia Pinkham data, in addition to the annual sales, we have corresponding observations on the annual expenditure on advertising the product – see Figure 1.1b. Aims of studying such data (in addition to those already stated for univariate time series) include the description and understanding of the ways in which the variables interact dynamically – for example, we would be interested to know if advertising effort (measured by the expenditure on it) affected sales at some time in the future (as we might hope). It may also happen that an unexpected relationship exists – current sales may affect future advertising (since higher sales may allow more revenue to be spent on advertising) as appears to be the case with the Lydia Pinkham data (see §A.1). Thus, any technique which we apply to explore the relationships between time series must be flexible enough to accommodate a wide range of dynamic structures. Once we have developed a satisfactory description of the system, we may wish to control one variable in order to influence the future values of another (particularly if the variables are industrial process parameters, one of which is crucial – such as a measure of the quality of the end product), or we may hope that a full description of the interrelationships among the variables will enable us to predict the future values with greater accuracy than would be possible when considering each variable singly.

1.1.2 Time series models

In this thesis we will address the problem of building models for multivariate time series data. These models may be used to summarise the dynamic relationships in some way (perhaps with a view to explaining the observed interactions) and/or to forecast future values, but we will be mostly concerned with methods of specifying "good" models for the data. For general multivariate systems, we will require a flexible class of models, several candidates of which may be thought to adequately describe the system. This flexibility necessarily leads to problems with choosing candidate models and restricting the parameters involved to a manageable number (since for multivariable systems we will usually require a large number of parameters to describe complex relationships). We will use models from the vector autoregressive moving–average (VARMA) class to be described in Chapter 2, which possesses the required flexibility. We discuss some other possible representations and further modelling considerations in §2.6.

1.2 Univariate ARMA models

The VARMA model of Chapter 2 is a multivariate generalisation of the univariate autoregressive moving–average (ARMA) model of Box and Jenkins (1976). We will introduce the general form of the univariate class together with the necessary

notation and some modelling details. The models are stochastic, since few observed phenomena will be entirely deterministic and although few time series are thought to be exactly generated by an ARMA process, it is hoped that some candidate models can be used to adequately describe observed series.

1.2.1 Model form

The ARMA model takes the form of a linear stochastic difference equation

$$z_t=\varphi_1 z_{t-1}+\ldots+\varphi_p z_{t-p}-\theta_1 a_{t-1}-\ldots-\theta_q a_{t-q}+a_t, \qquad [1.1]$$

where z_t is the value of our observed variable recorded at time t, a_t is a white noise process (each realisation a_t is independent of the others and they are identically distributed as $N(0,\sigma_a^2)$ random variables) and φ_i and θ_i are the parameters in the model. This is the general form of the ARMA model (of order (p,q)) which can also be written as

$$\varphi(B)z_t=\theta(B)a_t, \qquad [1.2]$$

where $\varphi(B)=1-\varphi_1 B-\ldots-\varphi_p B^p$ and $\theta(B)=1-\theta_1 B-\ldots-\theta_q B^q$ are polynomials in the backshift operator B (see notation). $\varphi(B)$ is the AR(p) part of the model and $\theta(B)$ the MA(q) part. A mean term μ or θ_0 may be included to account for the level of the process (see also §3.1.1(c)), but for simplicity we assume that the data z_t, has been mean–corrected. The form [1.2] is an approximation to the MA(∞) representation of a stationary stochastic process (see §2.1.1) by a ratio of polynomials. A stochastic process is said to be **second–order** stationary if its mean and variance are constant (over time) and the covariance between observations depends only on the time difference between them (Box and Jenkins (1976) for example, also define **strict** stationarity, for which second–order stationarity is sufficient under the normality assumption).

A wide range of different behaviour can be produced from the form [1.1] by varying the parameters – Figure 1.2a shows the series generated from the ARMA(1,0) (or AR(1)) process with $\varphi_1=0.9$, while Figure 1.2b shows that from the MA(1) process with $\theta_1=0.4$ (both simulations use $\sigma_a^2=1$) – so that this class of models should be sufficiently flexible to describe many observed series.

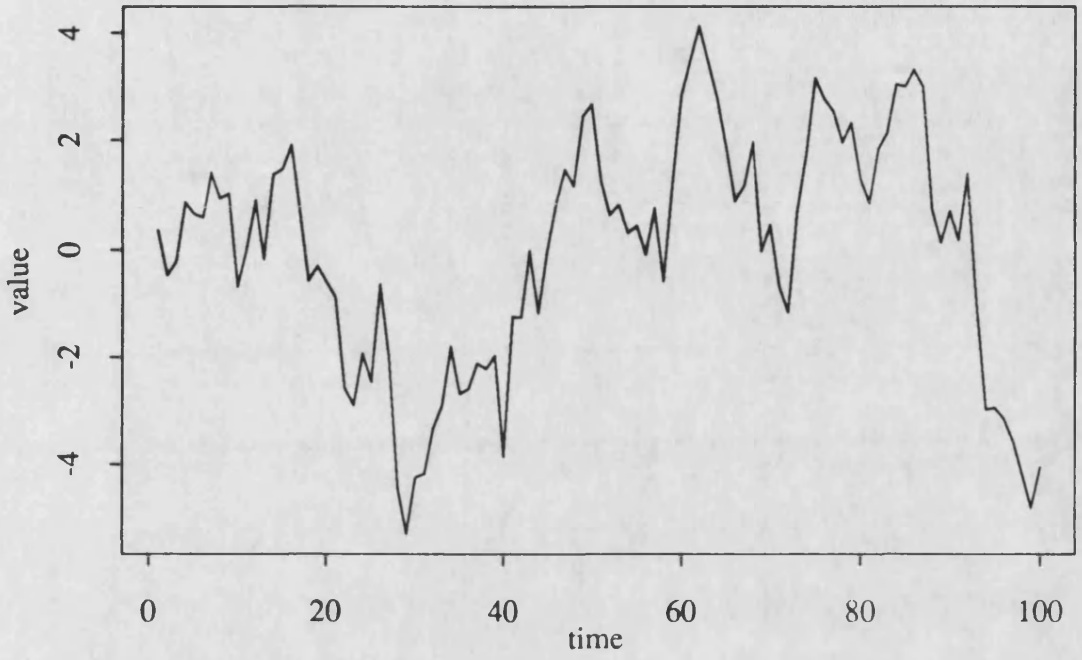
1.2.2 Model building

Box and Jenkins (1976, Figure 1.7) describe the iterative process of model building and highlight the following stages:

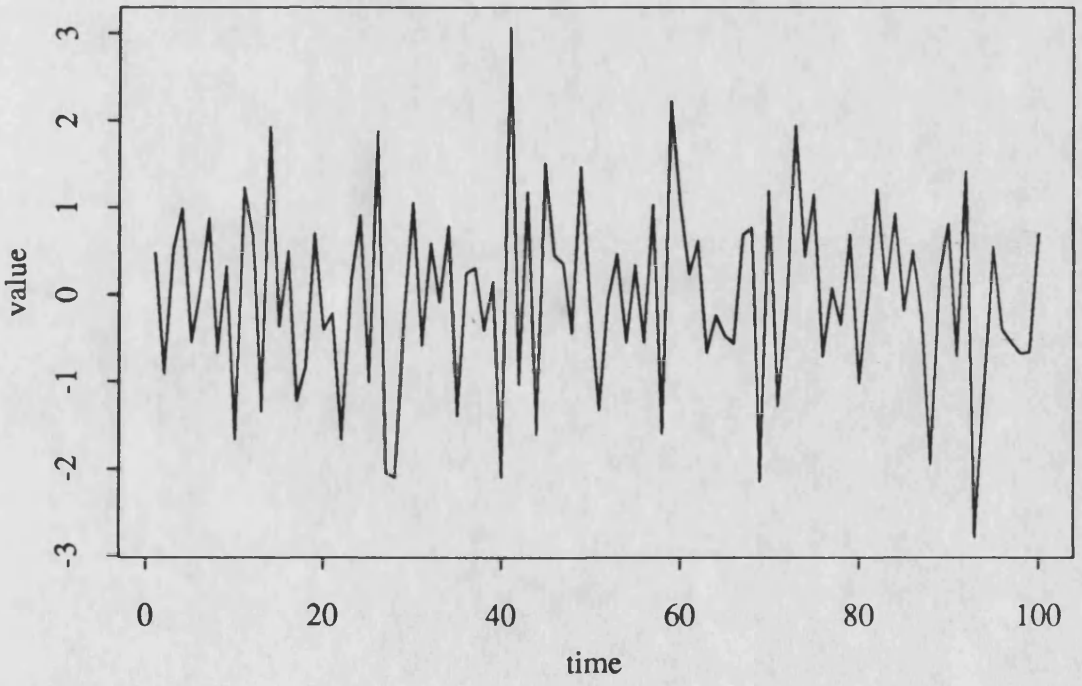
(a) Postulate a general class of models (the ARMA class in this case)

Figure 1.2

(a) AR(1) series



(b) MA(1) series



(b) Identify from this class particular model(s) to be tentatively entertained

(c) Estimate parameters in model(s)

(d) Check model adequacy – if model is not adequate then return to stage (b) (or perhaps even (a) if there is serious deficiency)

(e) Use the model – often for forecasting.

We will not discuss the building of univariate ARMA models in detail since we will present the equivalent multivariate methods in later chapters. Box and Jenkins (1976) and Wei (1990) for example, provide comprehensive coverage of ARMA model building with identification (stage (b)) based upon the univariate autocorrelation function (acf), defined to be

$$\rho_h = \frac{E[z_t z_{t+h}]}{\mathrm{var}(z_t)} = \frac{E[z_t z_{t-h}]}{\mathrm{var}(z_t)}, \quad h=0,1,2,... \qquad [1.3]$$

and the partial autocorrelation function (pacf), φ_{hh} which is defined to be the correlation between z_t and z_{t+h} after their mutual linear dependency on the intervening variables $z_{t+1},...,z_{t+h-1}$ has been removed. Wei (1990, section 2.5) defines the sample functions (e.g. $\hat{\varphi}_{hh}$ is given by the estimated coefficient $\hat{\varphi}_h$ when fitting an AR(h) model to the data) which are necessary to identify particular candidate ARMA models for given datasets. For an MA(q) process, $\rho_h=0$, $h=q+1,q+2,...$ and for an AR(p) process, $\varphi_{hh}=0$, $h=p+1,p+2,...$ and the theoretical correlation functions of mixed ARMA(p,q) models in general follow more complex patterns. As a general yardstick, the acf and pacf of a white noise process have approximate variance $1/N$ (where N is the number of observations available), so that for instance, $2/\sqrt{N}$ is often used as a guide to whether the individual coefficients in a correlation function are insignificant, signifying a cut–off which may suggest a particular model order (see e.g. Wei (1990, Chapter 6) for more details). We can then estimate the necessary parameters (φ_i, θ_i) in the chosen model (see e.g. Wei (1990, Chapter 7) for details of estimation procedures) and use various measures (often based upon the estimated residual series $\hat{a}_t$ – see e.g. Box and Jenkins (1976, Chapter 8) or Wei (1990, Chapter 7)) to assess the adequacy of the model – refining it with further/fewer parameters or a different form if necessary.

Box and Jenkins (1976) provide a thorough review of the details and results required to produce forecasts from univariate ARMA models and we present the equivalent multivariate results in §2.5.

1.3 Transfer function models

As a first step towards modelling the relationships between multivariate series Box and Jenkins (1976) describe the transfer function model for a (linear) system with a single input variable which is filtered in some way to produce a single output variable. This formulation assumes that we know the direction of causation, as will often be the case in for example industrial processes, but only occasionally in more general situations (although for example, certain price series may be affected by a global variable such as interest rates, but will themselves have only a very minor (feedback) effect on that variable). If the output does not affect the input of a system, the input series is said to be **exogenous.**

1.3.1 Model form

The general (single input, single output) transfer function (or **transfer function noise** = TFN) model takes the form

$$y_t=v(B)x_t+n_t \qquad [1.4]$$

where x_t is our (exogenous) input series, y_t the output, $v(B)=v_0+v_1B+v_2B^2+\ldots$ is the **transfer function** or filter and n_t is the noise series for the system which is independent of x_t, but **not** necessarily white noise (in general n_t is approximated by an ARMA form $\frac{\theta(B)a_t}{\varphi(B)}$, for some white noise series a_t). It can be seen that the univariate ARMA model [1.2] is a special case of [1.4] with $v(B)=0$ (i.e. there is no input series).

The coefficients v_i are called the impulse response weights since they measure the effect that a unit impulse input x_t (=1 at time t and 0 otherwise) has on y_{t+i}. To obtain an adequate representation, $v(B)$ must in general be an infinite order polynomial in B, so it is usually approximated by a ratio of finite order polynomials

$$v(B)=\frac{\omega(B)B^b}{\delta(B)} \qquad [1.5]$$

where $\omega(B)=\omega_0-\omega_1B-\ldots-\omega_sB^s$, $\delta(B)=1-\delta_1B-\ldots-\delta_rB^r$ and $b\geq 0$ is a lag between the impulse x_t affecting y_{t+b}, which may often be present in systems (the coefficients v_i, ω_i and δ_i can be found from one another by equating coefficients in the polynomials).

1.3.2 Model building

We will briefly outline the procedure used to build TFN models. The same iterative stages of identification, estimation and checking outlined in §1.2.2 for

ARMA models can be applied, where now we look for a possible lag b and choose tentative orders s and r for the rational approximation to the transfer function (the parameters ω_i and δ_i of which can be estimated in a similar way to the univariate case – see e.g. Wei (1990, Chapter 13)). The resulting noise series $\hat{n}_t$ can then be analysed in order to select a suitable ARMA model for it, allowing the full TFN model to be estimated and the iterative checking and refinement procedure carried out until a suitable representation is found.

To identify the transfer function $v(B)$, we calculate the cross correlation function (ccf) of the series x_t and y_t, given by

$$\rho_{xy}(h)=\frac{E[x_t y_{t+h}]}{\sigma_x \sigma_y}=\rho_{yx}(-h) \qquad [1.6]$$

(e.g. σ_x^2 is the variance of x_t). If x_t is white noise (so that $\rho_{xx}(h)=\rho_x(h)=0$ for $h\neq 0$) then we find that (e.g. Wei (1990, Chapter 13))

$$v_h=\frac{\sigma_y}{\sigma_x}\rho_{xy}(h) \qquad [1.7]$$

and so the impulse response weights can be directly calculated from the ccf **if** the input is white noise. To achieve this we usually **prewhiten** the input series x_t – that is we replace x_t with α_t where we assume that we have a suitable ARMA model for x_t, $\varphi_x(B)x_t=\theta_x(B)\alpha_t$. We then apply the same prewhitening filter to the output series to obtain $\beta_t=\frac{\varphi_x(B)}{\theta_x(B)}y_t$ and can thus build a TFN model

$$\beta_t=v(B)\alpha_t+\varepsilon_t, \qquad [1.8]$$

using the result [1.7] for the white noise input series α_t. A model in terms of x_t and y_t can be obtained from [1.8] by substituting in the forms for β_t and α_t. Box and Jenkins (1976, Chapter 11) and Wei (1990, Chapter 13) detail the methods used to choose orders s and r for the polynomials $\omega(B)$ and $\delta(B)$ and the lag b in the rational approximation to $v(B)$, based upon patterns in the sample function $\hat{v}_h$ of [1.7].

1.3.3 Forecasting with transfer function models

A TFN model can be used to forecast future values of the output series y_t, based upon the past of both y_t and the input series x_t. This is particularly useful if we have a non–zero lag b in the model representation [1.5], in which case x_t is said to be a **leading indicator** of y_t and the **known** values of past x_t can be used to improve the forecasts of y_t (when compared with those produced by a univariate ARMA model for y_t) – see also Ashley (1983) and §3.4.1. If the lag $b=0$, then a forecast of y_t is based upon forecast values of x_t (usually obtained from the

univariate ARMA model $\varphi_x(B)x_t=\theta_x(B)\alpha_t$) in which case we may not get superior forecasts of the output from the TFN model. Wei (1990, Chapter 13) gives the details of using TFN models for forecasting and Jenkins (1979) and Lin (1989) give examples of their use for forecasting.

1.3.4 Extensions to the model

It is possible to extend the TFN model of [1.4] to include multiple independent input series x_{it} (each with their own transfer function $v_i(B)$), or a simultaneous feedback model (with the rôles of x_t and y_t reversed) if x_t is not exogenous (see e.g. Jenkins (1979)). These extensions will make the analysis considerably more complicated and although the TFN models perform well in some situations (see e.g. Jenkins (1979) and Jenkins and McLeod (1982)) and can be straightforward to build, they cannot be easily used to describe more general (i.e. not input–output) relationships.

1.4 Outline of thesis

To model general multivariate time series relationships we will introduce the VARMA class and the usual details of model building in Chapter 2 (based upon the same iterative approach of §1.2.2). We also discuss the estimation of parameters and ways of checking the adequacy of the models – the procedures described will be illustrated with case studies in Chapter 6. Results necessary to use the models for forecasting will be presented in §2.5 and some other possible models and the situations when they may be preferable will be discussed in §2.6.

Chapter 3 considers some common problems with building VARMA models for time series data, including ways of handling data which deviate from the modelling assumptions of Chapter 2 and the problems of applying the usual model identification techniques. We also consider ways of comparing different VARMA models (when several possible models are suggested by some identification procedure) in order to choose between them. Some aspects of forecasting (including forecast comparison) and the effects on the models of transforming the variables are also discussed.

In Chapter 4 we consider the notion of co–integration among multivariate time series variables and the effects it has upon different models for the data. We present and extend the methods of testing for it, due to Engle and Granger (1987) and develop a strategy to employ when analysing co–integrated data in the context of VARMA modelling. We illustrate this with data simulated from suitable models and a case study using real data. Possible extensions to the techniques and some other approaches are also considered.

Chapter 5 describes and compares several methods of specifying parsimonious VARMA models (due to Tiao and Tsay (1989) and Tsay (1989)). We clarify some of the theoretical results and consider extensions to the techniques as well as exploring the relationships between the methods. Some case studies in Chapter 6 (using data described in the Appendix) allow us to compare the different methods with each other as well as with the usual model identification tools. We can then make recommendations on the use of the various procedures. We also consider the modelling implications of co–integration in Chapter 6.

Chapter 7 provides a summary of the results and a general strategy to employ when building VARMA models for multivariate time series data, together with some suggestions for further research.

Chapter 2. The VARMA model

In this chapter we will introduce the general form of the VARMA model for multivariate time series, including its origins, the different possible representations and the conditions which we require on the models. We also present the usual tools to identify a particular model from the class to represent an observed series (§2.2) and discuss the estimation of the parameters of a chosen model (§2.3) and the diagnostic checking procedures used to assess the adequacy of a fitted model (§2.4). VARMA models can be used to forecast the future values of the variables and we present the necessary results for this in §2.5. In §2.6 we discuss some other models for multivariable systems.

2.1 Model form

Vector autoregressive moving-average (VARMA) models are a natural multivariate generalisation of the univariate ARMA model described in §1.2. The model form is

$$\mathbf{z}_t=\Phi_1\mathbf{z}_{t-1}+\ldots+\Phi_p\mathbf{z}_{t-p}-\Theta_1\mathbf{a}_{t-1}-\ldots-\Theta_q\mathbf{a}_{t-q}+\mathbf{a}_t, \qquad [2.1]$$

where $\mathbf{z}_t=(z_{1t},\ldots,z_{kt})^T$ is a vector of observations at time t, $\mathbf{a}_t=(a_{1t},\ldots,a_{kt})^T$ is a realisation of an i.i.d. $N_k(\mathbf{0},\Sigma_a)$ random variable and Φ_i, Θ_i are $k\times k$ matrices of parameters. $\varphi_{ij}^{[h]}=[\Phi_h]_{ij}$ (the ij'th element of Φ_h) and $\theta_{ij}^{[h]}$ are measures of the effects that changes in series z_{jt} and residual series a_{jt} have on series $z_{i,t+h}$. It is assumed for simplicity throughout that $\mathbf{z}_t$ has been mean–corrected, although as in the univariate case, we may include a level term μ or Θ_0. This model is said to be a VARMA(p,q) model and can also be written as

$$\Phi(B)\mathbf{z}_t=\Theta(B)\mathbf{a}_t, \qquad [2.2]$$

where $\Phi(B)=I-\Phi_1B-\ldots-\Phi_pB^p$ and $\Theta(B)=I-\Theta_1B-\ldots-\Theta_qB^q$ are matrix polynomials in the backshift operator B (see notation).

As examples of the behaviour of VARMA processes we display Figure 2.1 – 2.1a and 2.1b show the two component series generated by the VAR(1) process

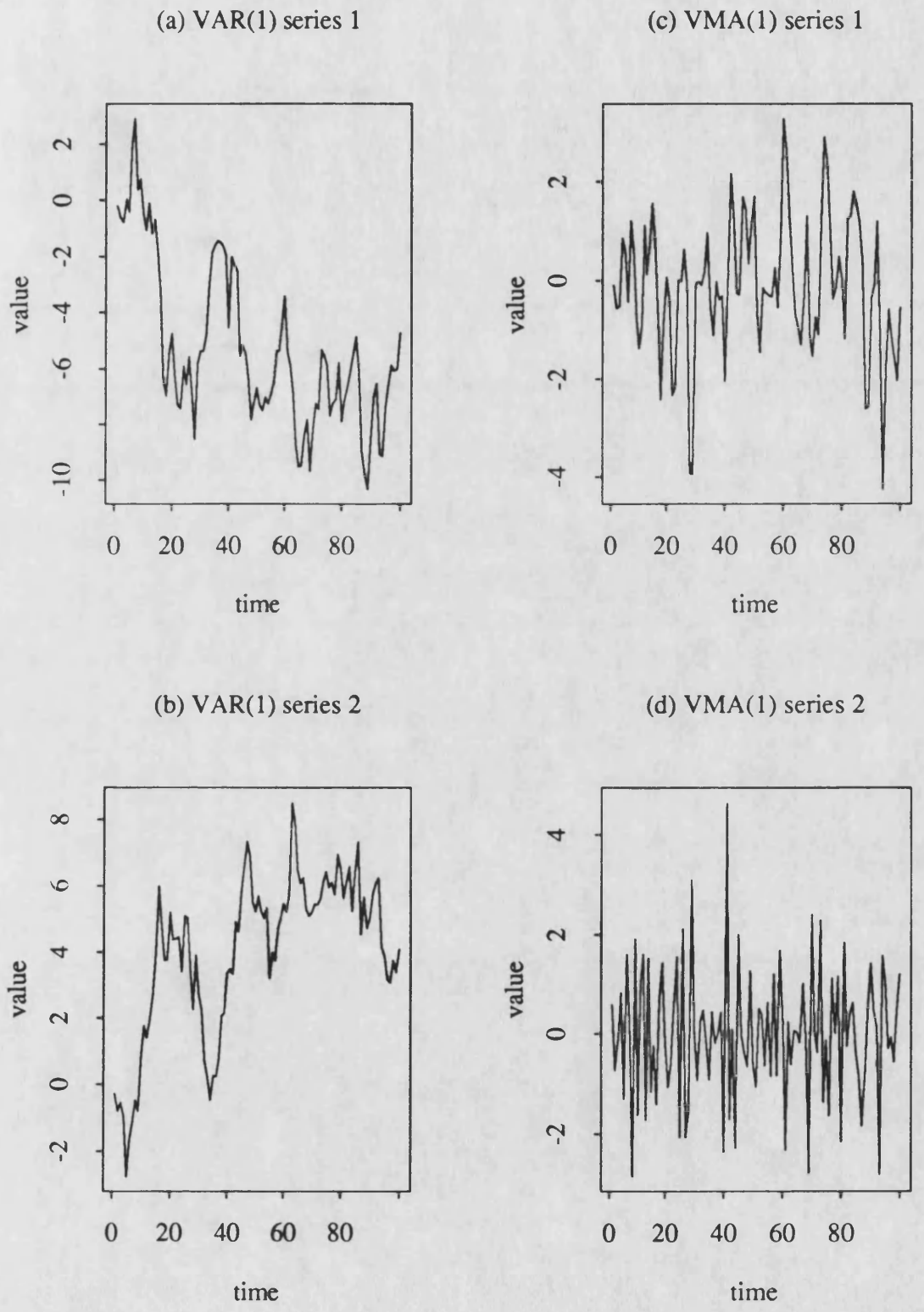
$$\mathbf{z}_t=\begin{bmatrix}0.60 & -0.50\\ 0.00 & 0.95\end{bmatrix}\mathbf{z}_{t-1}+\mathbf{a}_t, \quad \Sigma_a=\begin{bmatrix}1.00 & \\ 0.00 & 1.00\end{bmatrix} \qquad [2.3]$$

and 2.1c and 2.1d those from the VMA(1) process

$$\mathbf{z}_t=\mathbf{a}_t-\begin{bmatrix}-0.60 & 0.40\\ 0.20 & 0.80\end{bmatrix}\mathbf{a}_{t-1}, \quad \Sigma_a=\begin{bmatrix}1.00 & \\ 0.80 & 1.00\end{bmatrix}. \qquad [2.4]$$

These processes have very different characteristics – component z_{2t} of process [2.3] is exogenous and nearly non–stationary and it drives the component z_{1t}; for

Figure 2.1


(a) VAR(1) series 1
(c) VMA(1) series 1
(b) VAR(1) series 2
(d) VMA(1) series 2
value
time
0
20
40
60
80
-10
-6
-4
-2
0
2
4
6
8

process [2.4], the first component in some way smooths out the variation in both of the white noise processes – it is hoped that the VARMA model is sufficiently flexible to describe many observed relationships.

2.1.1 Origins

VARMA models appear to have first been studied by Quenouille (1968 – first edn. 1957) as special cases of the MA(∞) representation of a vector process. Hannan (1970) proves the Wold decomposition theorem for multivariate time series which states that if z_{it} are each stationary and purely non–deterministic and the collection $\{\mathbf{z}_t\}$ are jointly covariance stationary (i.e. $\text{cov}(z_{it},z_{j,t-h})$ is independent of t for all i,j,h) then $\mathbf{z}_t$ can always be represented uniquely by

$$\mathbf{z}_t=\Psi(B)\mathbf{a}_t, \qquad [2.5]$$

where $\mathbf{a}_t$ are i.i.d. with mean $\mathbf{0}$ and covariance matrix Σ_a, and $\Psi(B) = \sum_{i=0}^{\infty}\Psi_i B^i$ is an infinite order matrix polynomial in B. We usually assume that the conditions required on $\mathbf{z}_t$ hold – i.e. our data has been suitably transformed (but see §3.1). Usually we take $\Psi_0=I$ and Σ_a not necessarily diagonal, for convenience, but, for instance, Quenouille assumes that the white noise series a_{it} are **each** i.i.d., mean 0, variance σ^2 i.e.

$$\mathbf{z}_t=(\Psi_0+\Psi_1 B+\dots)\mathbf{b}_t, \quad E[\mathbf{b}_t\mathbf{b}_t^T]=\sigma^2 I.$$

This is equivalent to the form with $\Psi_0=I$ since Σ_a, being the covariance matrix of $\mathbf{a}_t$ must be positive definite in all but degenerate cases, and so we have that $\Sigma_a Q=Q\Lambda$, Λ is the diagonal matrix of eigenvalues (all positive) of Σ_a and Q is the matrix of eigenvectors (non–singular). So we have $\mathbf{b}_t=\sigma\sqrt{(\Lambda^{-1})}Q^T\mathbf{a}_t$ and $\Psi_0=Q\sqrt{(\Lambda)}/\sigma$.

We would obviously not attempt to build models such as [2.5] and, as in the univariate case, we may hope to approximate $\Psi(B)$ by the ratio of two finite order matrix polynomials

$$\Psi(B)\approx\Phi^{-1}(B)\Theta(B)$$

(Φ and Θ are as in [2.2], for suitable p and q and $|\Phi(z)|\neq 0$ for $|z|\leq 1$ and $|\Theta(z)|\neq 0$ for $|z|<1$ – Hannan and Diestler (1988)). Hannan and Diestler (1988) and Tiao and Tsay (1989 – reply to discussion) claim that representations such as [2.2] are appealing and have, in their experience, proven to be sufficiently flexible to be able to represent a wide variety of dynamic structures among the variables. The linearity inherent in the model and the normality assumption are also justifiable through experience and the tractability of the theory to which they give rise. We consider some other possible models and their relationships to the VARMA model in §2.6.

2.1.2 Representations

In the same way that we take $\Psi_0=I$ in [2.5], we also remove all contemporaneous effects in a model such as [2.2] to the covariance matrix of the noise term, but for example if we have

$$(\Phi_0+\Phi_1B+\ldots+\Phi_pB^p)\mathbf{z}_t=(\Theta_0+\Theta_1B+\ldots+\Theta_qB^q)\mathbf{b}_t,\ \mathrm{cov}(\mathbf{b}_t)=\Sigma_b, \quad [2.6a]$$

then equivalently

$$(I+\Phi_1'B+\ldots+\Phi_p'B^p)\mathbf{z}_t=(I+\Theta_1'B+\ldots+\Theta_q'B^q)\mathbf{a}_t, \quad [2.6b]$$

where $\Phi_i'=\Phi_0^{-1}\Phi_i$, $\Theta_i'=\Phi_0^{-1}\Theta_i(\Theta_0^{-1}\Phi_0)$, $\mathbf{a}_t=(\Phi_0^{-1}\Theta_0)\mathbf{b}_t$, $\Sigma_a=(\Phi_0^{-1}\Theta_0)\Sigma_b(\Phi_0^{-1}\Theta_0)^T$ (provided that $|\Phi_0|=|\Phi(0)|\neq 0$ and $|\Theta_0|=|\Theta(0)|\neq 0$, both of which certainly hold since we have that $|\Phi(z)|\neq 0$, $|z|\leq 1$ and $|\Psi_0|=|\Psi(0)|=|\Theta(0)|/|\Phi(0)|\neq 0\Rightarrow |\Theta(0)|\neq 0$). Thus all contemporaneous effects in model [2.6a] have been absorbed into the white noise term $\mathbf{a}_t$ of [2.6b].

The representation in [2.2] has become the most usual one, although Jenkins and Alavi (1981) and Granger and Newbold (1986) consider $\Phi(B)$ (and similarly $\Theta(B)$) as a single $k\times k$ matrix with each entry φ_{ij} being a scalar polynomial in B of order p_{ij}. The model is then of order (P,Q) where e.g. P is a $k\times k$ matrix with entries p_{ij}, so that the order of a VARMA model in the form [2.2] would now be $(\max(p_{ij}), \max(q_{ij}))$. This alternative form collects together all of the lagged effects of each series on every other one, but it becomes rather cumbersome as k becomes large since it writes the relationships as k simultaneous equations, making identification a larger problem (although it does have a close connection with the Scalar Component Models of §5.2.1) and it is rarely used in practice.

Seasonality could be incorporated into the form [2.1] by allowing intermediate parameter matrices (e.g. $\Phi_2,\ldots,\Phi_{s-1}$, for some seasonal period, s) to be zero. Jenkins and Alavi suggest that in many cases, seasonal behaviour will not have any between-series effects, so that e.g. Φ_s may be diagonal. They also suggest that parsimony may lead to the use of multiplicative models, where in [2.2] for example, $\Phi(B)=\Phi_1(B)\Phi_2(B^s)$, although in view of the problems encountered when trying to build even simple VARMA models, this multiplicative representation is perhaps too complex. We consider seasonality in later chapters.

Hannan and Diestler (1988) present a thorough account of ARMAX models. These are models such as [2.2] with a term $A(B)\mathbf{x}_t$ added onto the right hand side. $\mathbf{x}_t$ are variables **exogenous** (X for exogenous) to the system of interest, $\mathbf{z}_t$, but thought to have influence on it. $A(B)$ is a matrix polynomial describing the effect of these variables on $\mathbf{z}_t$. VARMA models are a special case of these so that Hannan and

Diestler (1988) contains many of the necessary theoretical results for VARMA modelling. Exogenous variables could be incorporated into the usual model [2.2] simply by extending the variables in $\mathbf{z}_t$ and constraining certain coefficients in the parameter matrices to be zero (see §2.6.1). Often though we do not have information available on the direction of causation, and want to explore it with a more general VARMA rather than an ARMAX model (see Granger and Newbold (1986, Chapter 7) for a discussion of causation).

2.1.3 Conditions on the representation [2.2]

In order to apply certain results (particularly for parameter estimation) to models in the form [2.2] we need to impose certain conditions on the representation.

Stationarity

If $|\Phi(z)| \neq 0$ for $|z| \leq 1$, the AR operator $\Phi(B)$ is said to be stationary and we can write

$$\mathbf{z}_t = \Phi^{-1}(B)\Theta(B)\mathbf{a}_t = \Psi(B)\mathbf{a}_t . \qquad [2.7]$$

This condition ensures that the expansion $\sum_{i=0}^{\infty} \Psi_i B^i$ converges for $|B| \leq 1$ (in that $\sum_{h=0}^{\infty} (\psi_{ij}^{[h]})^2 < \infty \ \forall i,j$ – Wei (1990)) and in particular that the covariance matrix of the (stationary) series $\mathbf{z}_t$ has finite elements. We will deal with departures from stationarity in §3.1, but for the moment assume that our VARMA models are specified so as to satisfy this condition. The elements $\psi_{ij}^{[h]}$ of the matrices Ψ_h describe the effect of the "random shock" $a_{j,t-h}$ on z_{it}, so that the matrices Ψ_h are sometimes called the impulse response function of the model.

Invertibility

If $|\Theta(z)| \neq 0$ for $|z| \leq 1$, the MA operator $\Theta(B)$ is said to be invertible and we can write

$$\Theta^{-1}(B)\Phi(B)\mathbf{z}_t = \Pi(B)\mathbf{z}_t = \mathbf{a}_t . \qquad [2.8]$$

This condition ensures that the expansion of $\Pi(B)$ (which is usually written as $I - \sum_{i=1}^{\infty} \Pi_i B^i$) converges for $|B| \leq 1$ ($\sum_{h=1}^{\infty} |\pi_{ij}^{[h]}| < \infty \ \forall i,j$ – Wei (1990)). Jenkins and Alavi (1981) build up the AR(∞) representation [2.8] from forecasting considerations – it is sensible to use weighted sums of past observations on **all** of the series as forecasts of the future values of each of them. The elements $\pi_{ij}^{[h]}$ of Π_h describe the effect of the past observation $z_{j,t-h}$ on z_{it} and the invertibility condition then ensures that the forecast weight functions $\pi_{ij}^{[h]}$ die out for increasing

h and hence that the forecasts depend less upon the more remote than the more recent past.

Uniqueness

Hannan (1969) gives conditions for the representation [2.2] to be unique (identifiable):

A. $|\Phi(z)| \neq 0$ for $|z| \leq 1$, $|\Theta(z)| \neq 0$ for $|z| \leq 1$ (stationarity and invertibility).

B. If $\Phi(B)=C(B)H(B)$ and $\Theta(B)=C(B)K(B)$, for some matrix polynomials $C(B)$, $H(B)$ and $K(B)$ then $|C(B)|$ must be a constant (i.e. Φ and Θ are left co–prime).

C. The null spaces of Φ_p^T and Θ_q^T have null intersection. This condition is equivalent to requiring that $(\Phi_p - \Theta_q)$ has no left eigenvectors associated with zero eigenvalues (i.e. is of full rank), or that the matrix $[\Phi_p : \Theta_q]$ (see notation) is of full rank (Granger and Newbold (1986) and Wei (1990)).

Given data generated from a model in the form of [2.2] which satisfies these conditions (A, B and C), we can then uniquely identify the model using some procedure. Wei (1990) quotes Hannan as suggesting that we could choose those models with minimal MA order, q, then of these, one with minimal AR order, p. Of course, we will rarely have data which exactly follows any particular VARMA model, but need to choose the model which best describes the observed relationships.

2.2 Order identification

The first stage in building a VARMA model such as [2.2] for real data $\mathbf{z}_t$ (which has been suitably transformed to satisfy any conditions we may need – see §2.1.1 and §3.1) is to choose orders p and q for the matrix polynomials $\Phi(B)$ and $\Theta(B)$. Obviously, we would like these to be as small as possible and we may hope to further minimise the number of parameters involved with good initial estimates of possibly zero coefficients in the parameter matrices. Generalising the univariate Box–Jenkins approach to model building means that we study the behaviour of the correlation structure of some known models in order to compare the observed behaviour of our data with this. The autocorrelation and partial autocorrelation functions for univariate time series are well–known and useful tools for identifying univariate ARMA models (§1.2.2). We now discuss the multivariate generalisations of these functions.

2.2.1 MA models – cross–correlations

The cross–covariance function (ccv, also called the covariance matrix function) of a stationary process $\mathbf{z}_t$ (which is assumed to have zero mean) is defined to be

$$\Gamma(h)=\text{cov}(\mathbf{z}_t,\mathbf{z}_{t+h})=E[\mathbf{z}_t\mathbf{z}_{t+h}^T]=\{\gamma_{ij}^{[h]}\},\quad h=0,\pm1,\pm2,\ldots \qquad [2.9]$$

(or equivalently $\Gamma(h)=E[\mathbf{z}_{t-h}\mathbf{z}_t^T]$), where $\gamma_{ij}^{[h]}=E[z_{it}z_{j,t+h}]=\gamma_{ji}^{[-h]}$. (Some authors define the cross–covariance function as $\Gamma^T(h)$, perhaps for convenience in their particular application, but [2.9] appears to be the most common definition.) The matrix elements can be estimated for observed data by

$$\hat{\gamma}_{ij}^{[h]}=\frac{1}{N}\sum_{t=1}^{N-h}z_{it}z_{j,t+h},\quad i,j=1,\ldots,k \qquad [2.10]$$

(N is the number of observations available on each series, z_{it}). The cross–correlation function (ccf, also called the correlation matrix function) is then defined to be

$$\rho(h)=\{\rho_{ij}^{[h]}\},\ \rho_{ij}^{[h]}=\frac{\gamma_{ij}^{[h]}}{(\gamma_{ii}^{[0]}\gamma_{jj}^{[0]})^{\frac{1}{2}}},\quad h=0,\pm1,\pm2,\ldots \qquad [2.11]$$

and each coefficient $\rho_{ij}^{[h]}$ can be estimated by replacing $\gamma_{ij}^{[h]}$ in [2.11] with their sample functions from [2.10]. $\hat{\rho}_{ij}^{[h]}$ measures the (linear) correlation between z_{it} and $z_{j,t+h}$. $\Gamma(h)$ and $\rho(h)$ are not necessarily symmetric for $h\neq0$ and so do not share the reversal property of the univariate functions (note that the above definitions correspond exactly to the univariate functions in the case of $\mathbf{z}_t=z_t$, scalar). However, it easy to see that

$$\Gamma(-h)=\Gamma^T(h)\ \text{ and }\ \rho(-h)=\rho^T(h),\quad h=0,\pm1,\pm2,\ldots$$

For an MA(q) process,

$$\mathbf{z}_t=\mathbf{a}_t-\Theta_1\mathbf{a}_{t-1}-\ldots-\Theta_q\mathbf{a}_{t-q}=-\sum_{i=0}^{q}\Theta_i\mathbf{a}_{t-i}\quad(\Theta_0=-I), \qquad [2.12]$$

the theoretical cross–covariance function has the form

$$\Gamma(h)=E\left[\left[\sum_{i=0}^{q}\Theta_i\mathbf{a}_{t-i}\right]\left[\sum_{j=0}^{q}\mathbf{a}_{t+h-j}^T\Theta_j^T\right]\right],\quad h=0,\pm1,\pm2,\ldots$$

The expectation is non–zero only when $t-i=t+h-j$ or $j=i+h$ (since $\mathbf{a}_t$ is white noise), giving us

$$\Gamma(h)=\begin{cases}\sum_{i=0}^{q-h}\Theta_i\Sigma_a\Theta_{i+h}^T & h=0,\ldots,q\\ 0 & h>q\end{cases},\quad \Gamma(-h)=\Gamma^T(h) \qquad [2.13]$$

when we take the expectation through the sum and matrix product. The cut–offs

in this function (and also therefore in the cross–correlation function) give a way of identifying the order, q, of a moving average process in the same way as in the univariate case. Hannan (1970) shows that $\hat{\rho}(h)$ is a consistent estimator of $\rho(h)$ which is asymptotically normally distributed. Wei (1990) gives expressions for the variance and covariance of the sample cross–correlations, $\hat{\rho}_{ij}^{[h]}$, which depend on the underlying model and shows that if $\rho_{ij}^{[h]}=0$ for $|h|$ greater than some q (i.e. series i and j are uncorrelated after some lag q) then

$$\text{var}(\hat{\rho}_{ij}^{[h]}) \approx \frac{1}{N-h}\left[1+2\sum_{s=1}^{q}\rho_{ii}^{[s]}\rho_{jj}^{[s]}\right], \quad |h|>q. \qquad [2.14]$$

To check for cut–offs in the sample ccf of a series, we can compare each coefficient $\hat{\rho}_{ij}^{[h]}$ with $\pm 1.96\sqrt{}$(var from [2.14]). This is cumbersome to use and since we usually hope to have a large number of observations and also that low order MA models will be adequate, Tiao and Box (1981) suggest that the value $2/\sqrt{N}$ is a suitable practical yardstick with which to replace $1.96\sqrt{}$(var) (this assumes that the sum of the autocorrelations in [2.14] is small, which is exactly true if $\mathbf{z}_t$ is white noise, but may underestimate the variance and lead to overparameterisation if the series are highly autocorrelated). We have found this approximation to be satisfactory, but for comparison calculate the more accurate variances from [2.14] in some of the examples in Chapter 6.

With Jenkins and Alavi's (1981) representation (§2.1.2) it is natural to look at the individual cross–correlation coefficients $\hat{\rho}_{ij}^{[h]}$ in order to suggest orders q_{ij} for the **individual** MA models. With the more common representation [2.2] it is usual to consider the correlation matrices $\hat{\rho}(h)$ ([2.11]) as indicators of an **overall** MA order q. To simplify this, Tiao and Box (1981) suggest the following device for displaying sample cross–correlation matrices:

Replace the entries, $\hat{\rho}_{ij}^{[h]}$ of the matrices by

$$\begin{array}{lllll} \text{"+"} & \text{if} & \hat{\rho}_{ij}^{[h]} & > & 2/\sqrt{N} \\ \text{"–"} & \text{if} & \hat{\rho}_{ij}^{[h]} & < & -2/\sqrt{N} \\ \text{"."} & \text{if} & |\hat{\rho}_{ij}^{[h]}| & < & 2/\sqrt{N} \end{array} \qquad [2.15]$$

Cut–offs can then be spotted more easily. Examples of this device in practice are given in Chapter 6.

2.2.2 AR models – partial correlation

For an AR(p) process,

$$\mathbf{z}_t = \Phi_1\mathbf{z}_{t-1}+\ldots+\Phi_p\mathbf{z}_{t-p}+\mathbf{a}_t, \qquad [2.16]$$

we can obtain recursions for the theoretical cross–covariance function, namely

$$\Gamma(0)=\Gamma(-1)\Phi_1^T+\ldots+\Gamma(-p)\Phi_p^T+\Sigma_a \quad \text{and} \qquad [2.17a]$$

$$\Gamma(h)=\Gamma(h-1)\Phi_1^T+\ldots+\Gamma(h-p)\Phi_p^T, \quad h\geq 1, \quad \Gamma(-h)=\Gamma^T(h) \qquad [2.17b]$$

so that this function simply decays for (stationary) AR models and does not give any indication of possible AR model orders. For this, we need something analogous to the univariate partial autocorrelation function (pacf) of §1.2.2, possible generalisations of which we now consider.

In the scalar case the pacf at lag h is defined to be the correlation between z_t and z_{t+h} after their mutual linear dependency on the intervening $z_{t+1},\ldots,z_{t+h-1}$ has been removed. Analogously to the univariate case, Tiao and Box (1981) define the (multivariate) **partial autoregression** matrix (parm) at lag h, $\mathcal{P}(h)$ to be the coefficient matrix Φ_h of a VAR(h) model fitted to the data (by a multivariate generalisation of the Yule–Walker equations – see e.g. Wei (1990)). This matrix in some way measures only the effect that $\mathbf{z}_t$ has on $\mathbf{z}_{t+h}$ neglecting that due to mutual correlation with intervening $\mathbf{z}$'s and if the process is generated by an AR(p) model ([2.16]) then $\mathcal{P}(h)=0$, $h>p$. This definition coincides with the univariate pacf in the case of scalar z_t, but in the multivariate case the elements of $\mathcal{P}(h)$ are not true correlation coefficients (they are not scale–invariant – see e.g. Table 6.4b in §6.2). Tiao and Box derive the necessary sample statistic to check for cut–offs in the sample function $\hat{\mathcal{P}}(h)$, which suggest possible AR model orders. This has the form

$$M(p)= -(N-\tfrac{1}{2}-pk)\ln U \qquad [2.18]$$

where U is the likelihood ratio $\dfrac{|S(p)|}{|S(p-1)|}$, $S(p)=\hat{\Sigma}_a$ from a VAR(p) model. Asymptotically M has a $\chi^2_{k^2}$ distribution, so that we reject the null hypothesis that $\Phi_p=\mathcal{P}(p)=0$ if U is small or $M(p)$ is large.

Heyse and Wei (1985a) mention Ansley and Newbold's multivariate **partial autocorrelation matrix** (pccf) $Q(h)$ which is another attempt to generalise the pacf. It shares the cut–off property of $\mathcal{P}(h)$ and coincides with the pacf for scalar series, however the elements of $Q(h)$ are again not true correlation coefficients and, in the multivariate case, $Q(1)\neq\rho(1)$ (neither does $\mathcal{P}(1)$), as it is for the univariate function. To overcome these discrepancies, Heyse and Wei (1985a) proposed their partial lag autocorrelation function (plcm).

Partial lag autocorrelation function

As a proper multivariate extension of the univariate pacf, Heyse and Wei (1985a) (see also Wei (1990)) derive the correlation between $\mathbf{z}_t$ and $\mathbf{z}_{t+h}$ after removing the linear dependency on the intervening vectors $\mathbf{z}_{t+1},\ldots,\mathbf{z}_{t+h-1}$. This is called the

partial lag autocorrelation function (plcm – partial lag correlation matrix), $P(h)$ and is defined to be the correlation between the residual vectors $\mathbf{u}_{t+h}$ and $\mathbf{v}_t$ from the multivariate linear regressions ($h \geq 2$)

$$\mathbf{z}_{t+h} = \alpha_1 \mathbf{z}_{t+h-1} + \ldots + \alpha_{h-1} \mathbf{z}_{t+1} + \mathbf{u}_{t+h} \qquad [2.19a]$$

$$\text{and } \mathbf{z}_t = \beta_1 \mathbf{z}_{t+1} + \ldots + \beta_{h-1} \mathbf{z}_{t+h-1} + \mathbf{v}_t. \qquad [2.19b]$$

(for $h=0,1$ there are no intervening vectors and $\mathbf{u}_{t+h}=\mathbf{z}_{t+h}$, $\mathbf{v}_t=\mathbf{z}_t$). $\mathbf{u}_{t+h}$ is the residual after any correlation between $\mathbf{z}_{t+h}$ and $\{\mathbf{z}_{t+h-1},\ldots,\mathbf{z}_{t+1}\}$ has been removed by the regression and $\mathbf{v}_t$ the residual after the correlation between $\mathbf{z}_t$ and $\{\mathbf{z}_{t+1},\ldots,\mathbf{z}_{t+h-1}\}$ has been removed. Heyse and Wei (1985a) show that this function has the required cut–off property for AR processes, its coefficients are true correlations, it coincides with the pacf in the scalar case and $P(1)=\rho(1)$ in the vector case. They also define the following in order to show the relationships between the various partial correlation functions: for $h \geq 2$ let

$$A=A(h)=\begin{bmatrix} \Gamma(0) & \cdots & \Gamma(2-h) \\ \vdots & \ddots & \vdots \\ \Gamma(h-2) & \cdots & \Gamma(0) \end{bmatrix},\ b=b(h)=\begin{bmatrix} \Gamma(1-h) \\ \vdots \\ \Gamma(-1) \end{bmatrix},\ c=c(h)=\begin{bmatrix} \Gamma(1) \\ \vdots \\ \Gamma(h-1) \end{bmatrix} \qquad [2.20]$$

and $A(1)=b(1)=c(1)=0$. Now define (dropping the index h for clarity)

$$\begin{aligned} V_{\mathbf{vu}}(h) &= \Gamma(h) - b^T A^{-1} c, \\ V_{\mathbf{u}}(h) &= \Gamma(0) - c^T A^{-1} c \text{ and} \\ V_{\mathbf{v}}(h) &= \Gamma(0) - b^T A^{-1} b \end{aligned} \qquad [2.21]$$

(which can be shown to be the variances ($V_{\mathbf{u}}$ and $V_{\mathbf{v}}$) and covariance ($V_{\mathbf{vu}}$) of the residuals $\mathbf{u}_{t+h}$ and $\mathbf{v}_t$ in [2.19] when the coefficient matrices α_i, β_i have been estimated by least squares from the linear regression (i.e. to minimise $E[|\mathbf{u}_{t+h}|^2]$ and $E[|\mathbf{v}_t|^2]$) – see e.g. Wei (1990, §14.5.3)). In addition let

$$W_{\mathbf{u}}^2(h)=V_{\mathbf{u}}(h),\ W_{\mathbf{v}}^2(h)=V_{\mathbf{v}}(h) \text{ (matrix square roots)}$$

$$D_{\mathbf{u}}(h)=\text{diagonal},\ [D_{\mathbf{u}}(h)]_{ii}=\sqrt{[V_{\mathbf{u}}(h)]_{ii}},\ D_{\mathbf{v}}(h) \text{ similarly.}$$

Heyse and Wei have shown that for $h \geq 1$

$$\begin{aligned} &\text{(parm)} \quad \mathcal{P}(h) = V_{\mathbf{vu}}^T V_{\mathbf{v}}^{-1} \\ &\text{(pccf)} \quad Q(h) = W_{\mathbf{u}}^{-1} V_{\mathbf{vu}}^T W_{\mathbf{v}}^{-1} \\ &\text{(plcm)} \quad P(h) = D_{\mathbf{v}}^{-1} V_{\mathbf{vu}} D_{\mathbf{u}}^{-1} \end{aligned} \qquad [2.22]$$

demonstrating that each "partial" function is a scaled version of the covariance between the residuals $\mathbf{u}_{t+h}$ and $\mathbf{v}_t$, but only the plcm correctly scales the matrix – in particular $[P(h)]_{ij}=\dfrac{[V_{\mathbf{vu}}]_{ij}}{\sqrt{[V_{\mathbf{v}}]_{ii}}\sqrt{[V_{\mathbf{u}}]_{jj}}}$. Wei (1990, p361.) for example gives a

recursive procedure for calculating $P(h)$. The sample plcm can be found by replacing $\Gamma(h)$ with their sample estimates in the equations [2.20], [2.21] and [2.22] above (or in the recursions). Under the null hypothesis that $\mathbf{z}_t$ follows a VAR($h-1$) model, the residual series $\mathbf{u}_{t+h}$ and $\mathbf{v}_t$ in [2.19] are uncorrelated (for $h\neq 0$) and both are white noise series. Wei (1990) shows that the elements $\hat{p}_{ij}^{[h]}=[\hat{P}(h)]_{ij}$ are then asymptotically i.i.d. $N(0,1/N)$. We can use Tiao and Box's device [2.15] to replace the elements of $\hat{P}(h)$ in order to spot cut–offs. In addition $N[\hat{p}_{ij}^{[h]}]^2$ are asymptotically i.i.d. χ_1^2, so that

$$X(h)=N\sum_{i=1}^{k}\sum_{j=1}^{k}[\hat{p}_{ij}^{[h]}]^2 \qquad [2.23]$$

is asymptotically $\chi_{k^2}^2$ distributed and can be used to detect cut–offs and so determine a possible AR order (we reject the hypothesis that $P(h)=0$ if $X(h)>\chi_{\alpha,k^2}^2$ for some chosen significance level α).

2.2.3 ARMA models

For mixed VARMA(p,q) processes, the first q cross–covariance matrices follow no fixed pattern (although, of course the theoretical function can be derived, but will not be presented here since it is of little immediate use), then for $h>q$ they satisfy the same recurrence relation as a pure VAR(p) model (see e.g. Jenkins and Alavi (1981)). Similarly the plcms decay and give no indication of model orders. Usually we try to identify a possible AR or MA order and then consider the residuals from fitting such a model although Tiao and Box (1981) point out that this may lead to incorrect identification due to biases in the estimates of the coefficient matrices Φ_i, when $q>0$.

q–conditioned parm (qparm)

To help identify mixed (ARMA) models, Jenkins and Alavi define the q–conditioned partial autoregression matrix. This assumes, stepwise, that the order of the MA polynomial is $q=1,2,..$ and for each of these uses the multivariate Yule–Walker equations to calculate $\mathcal{P}(h,q)$ (the parm). When the correct MA order q has been reached, $\mathcal{P}(h,q)$ will be zero for $h>p$, so that it first chooses an MA order and then an AR. The sampling properties of $\mathcal{P}(h,q)$ do not appear to have been studied and looking for cut–offs in the sample function is likely to be difficult and certainly very time consuming given the nature of the search. The function is analogous to the parm $\mathcal{P}(h)$ and is thus not properly normalised, as discussed in §2.2.2.

Extended sample cross–correlations (ESCC)

In order to overcome the bias in the estimates when estimating a VAR model for a VARMA process ($q>0$), Tiao and Tsay (1983) proposed their ESCC method for detecting the order of mixed models. They first show how consistent estimates of the parameter matrices Φ_i of the AR part can be obtained from recursively fitting AR(m),...,AR($m+j$) models, by least squares where m= a possible AR order and j= a possible MA order. These estimates are given by

$$\hat{\Phi}^{(j)}_{i(m)}=\hat{\Phi}^{(j-1)}_{i(m+1)}-\hat{\Phi}^{(j-1)}_{m+1(m+1)}\left[\hat{\Phi}^{(j-1)}_{m(m)}\right]^{-1}\hat{\Phi}^{(j-1)}_{i-1(m)},\quad i=1,\ldots,m$$

where $\hat{\Phi}^{(j)}_{i(m)}$ is the estimate of the matrix Φ_i of a VARMA(m,j) model for the process ($\Phi^{(j-1)}_{0(m)}=-I$). They then define the ESCC matrix to be $\hat{\rho}_{(m)}(j)$= the lag j sample cross–correlation matrix of the residuals from fitting the VAR(m) part obtained from using the above estimates of the parameter matrices. Tiao and Tsay show that this function has a cut–off property that $\hat{\rho}_{(m)}(j)\xrightarrow{P}0$ when $m=p$ (the true AR order) and $j>q$ (where $\xrightarrow{P}$ denotes convergence in probability – see notation), which would help to identify mixed models. They suggest variances for the elements of these matrices which allow such cut–offs to be spotted, but as with the q–conditioned parm's above, the ESCC matrices are probably too expensive to obtain for routine identification. Note that the ESCC method first chooses an AR order and then an MA.

While devices such as the qparm and ESCCs may aid identification of mixed models, the ccf and plcm have become the most usual tools for identifying a VARMA model order and for instance, the package MTS (Reilly (1987)) bases its identification procedure purely on these functions. Examples of such identification can be found in Chapter 6 and we will further discuss the problems associated with it in §3.2. We consider some other procedures which can be used to identify possible VARMA model orders in Chapter 5.

2.3 Parameter estimation

Once we have selected suitable orders p and q for a VARMA model, we then need to estimate the $(p+q)k^2$ parameters in the matrices Φ_i,Θ_i in [2.2] and the $\frac{1}{2}k(k+1)$ in Σ_a. In this section we discuss the techniques commonly used to do this.

2.3.1 Initial estimates

We might hope to have available from our preliminary identification tools, at least rough estimates of the values of some of the parameters in the suggested model in

order to perhaps impose some constraints on the model or as starting points for better estimation routines. For AR(p) models we can assemble the covariance matrix recursions [2.17b], $h=1,\ldots,p$ into a matrix form and using the sample estimates of $\Gamma(h)$, solve to obtain estimates of $\Phi_1,\ldots,\Phi_p$ in a multivariate generalisation of the Yule–Walker equations. Jenkins and Alavi (1981) claim that these provide good approximations to the maximum likelihood estimates. This procedure can also be applied in a slightly different form to obtain the AR coefficients of ARMA models (Shea (1987)), or the recursive method developed by Tiao and Tsay (1983) for their ESCC's could be used. For MA(q) models we could iteratively solve the set of equations given by [2.13] $h=0,\ldots,q$ (replacing $\Gamma(h)$ by their sample values and setting $\Theta_i=0$, $\Sigma_a=\Gamma(0)$, say, initially) for $\Theta_1,\ldots,\Theta_q$ and Σ_a. However, Jenkins and Alavi state that the convergence of this method is not satisfactory. The same would be true if it were to be used on mixed models. Shea (1987) presents a non–iterative method for ARMA models using a multivariate generalisation of Chatfield's (1979) inverse autocorrelations which allow the AR coefficients of the "inverse" model (which are the MA coefficients we require) to be calculated by the Yule–Walker equations. However, Shea points out that these estimates may sometimes not be satisfactory.

2.3.2 Conditional likelihood

Tunnicliffe Wilson (1973) shows that conditional on the initial values $\mathbf{z}_0,\ldots,\mathbf{z}_{1-p}$ being fixed (i.e. using the first p observations to initialise the routine) and $\mathbf{a}_0=\ldots=\mathbf{a}_{1-q}=\mathbf{0}$, the log of the likelihood function of a model in the form of [2.2] may be written as

$$L(\beta,\Sigma_a)= -\tfrac{1}{2}kN\ln(2\pi)-\tfrac{1}{2}N\ln|\Sigma_a|-\tfrac{1}{2}\sum_{t=1}^{N}\mathbf{a}_t^T\Sigma_a^{-1}\mathbf{a}_t \qquad [2.24]$$

where β is a parameter vector containing all of the coefficients from the Φ_i and Θ_i matrices (N is now the number of observations less the p we have used as initial values for the routine). Given some initial estimates of the parameters, we can recursively generate estimates of the residual series $\mathbf{a}_t$ by inverting the model [2.2]:

$$\mathbf{a}_t=\Theta^{-1}(B)\Phi(B)\mathbf{z}_t. \qquad [2.25]$$

In order to maximise $L(\beta,\Sigma_a)$, Tunnicliffe Wilson follows a two stage conditional estimation procedure which involves iteratively calculating $\hat{\mathbf{a}}_t$ and so $\hat{\Sigma}_a$, from [2.25] using the current parameter estimates. These parameter estimates can then be updated by differentiating [2.24] with respect to each parameter in turn and linearising the resulting equations about a point in order to approximately solve them. This is repeated until convergence and is shown to have suitable

convergence properties. Jenkins and Alavi claim that this procedure works quite well provided that $|\Theta(B)|$ does not have any zeroes close to the unit circle and the series are "not too short" (remembering that we lose some values in starting up the procedure). They mention a back–forecasting approximation to the likelihood function (where the initial values are forecast from the "backward" model, which replaces B with F such that $F\mathbf{z}_t=\mathbf{z}_{t+1}$, in the matrix polynomials) which they claim works well – see e.g. Wei (1990, Chapter 7) for details of backcasting in the univariate case.

2.3.3 Conditional least squares

Spliid (1983) presents an algorithm for estimating the parameters of a model such as [2.2] (maybe with exogenous variables – see §2.1.2) either conditional on p initial values or using back–forecasting to estimate them. He reformulates the estimation problem as a least squares regression which is solved iteratively and shown to converge, and derives the distributional properties of the estimators. No starting values are required since these are obtained within the procedure by initially fitting an AR($p+q$) model as an approximation to the AR(∞) representation ([2.8]). Spliid shows that the algorithm demands considerably less computation than a gradient method (such as that in §2.3.2) and illustrates this with an example. While the estimates are reasonably close to those obtained from a conditional likelihood procedure, they probably serve best as good initial estimates for a more efficient routine. Spliid's procedure is implemented (in conditional, not backcasting, form) in the package MTS (Reilly (1987)).

2.3.4 Exact likelihood

Hillmer and Tiao (1979) studied the exact likelihood function for VARMA models, the use of which reduces the bias in the estimates of the MA parameters when the zeroes of $|\Theta(B)|$ are close to the unit circle. Ansley and Kohn (1983) rewrite the VARMA model [2.1] in a state space form (see §2.6.2), the likelihood for which is easy to obtain and maximise using the Kalman filter. Shea (1987) also uses this approach, but presents a slightly different recursive algorithm which he claims to be "much" quicker than that of Ansley and Kohn, except in the common cases of AR(1), AR(2) and ARMA(2,1) models! This algorithm will benefit from good initial estimates of the parameters and is implemented in NAG routine g13dcf (Numerical Algorithms Group (1988), also Shea (1989) – see Chapter 6 for some examples of its use).

It is generally accepted that exact likelihood estimates are the "best" for all but a few peculiar cases, however they are comparatively expensive to obtain, so that some of the poorer procedures can be used in the preliminary stages of modelling

and to provide good starting points for the exact likelihood procedures when better estimates are required.

2.4 Diagnostic checking

Once the parameters of a possible model have been estimated, we would then check to see if the model appears to be adequate and if not, how it could be improved. The usual way to do this is to examine the residuals of the model. We should first ensure that all of the parameters are significant – the estimation routines also supply variances (and covariances) of the parameter estimates which enable insignificant parameters to be deleted. The model can then be re–estimated without them. The converse of this is to check whether we require additional parameters, perhaps by fitting extra ones (i.e. more terms in the polynomials) and seeing if any of them are significant. Of course, in the vector case, each increment in order adds k^2 correlated parameters, so that this overfitting is not as straightforward as in the univariate case.

Useful diagnostic checks on the residuals include their individual plots vs. time or other variables to reveal any structure or deviation from normality (as in the univariate case – see e.g. Brockwell and Davis (1987) and Box and Jenkins (1976)). Control limits of $\pm 2\sqrt{[\hat{\Sigma}_a]_{ii}}$ on the time plots will highlight any outliers, although Jenkins and Alavi suggest plotting the residuals transformed so as to make them uncorrelated (i.e. use $\hat{\mathbf{a}}_t^* = Q^T \hat{\mathbf{a}}_t$ where Q is the matrix of eigenvectors of $\hat{\Sigma}_a$, i.e. $\hat{\Sigma}_a Q = Q\Lambda$ where Λ is the diagonal matrix of eigenvalues, λ_i, of $\hat{\Sigma}_a$). Control limits of $\pm 2\sqrt{\lambda_i}$ can then be used **independently** for each series. Any suspected outliers may need correcting for (see §3.1.3) before the analysis continues. The correlation structure of the residuals could also be analysed using the techniques of §2.2 to see if they really do appear to be white noise and if not, the model for the data could be reformulated when combined with a possible VARMA model for the residuals, although this combination may be difficult. Hosking (1980) presents an overall (portmanteau) test based on the first s sample cross–covariances of the residuals, but Tiao and Box (1981), among others, warn against using such overall tests without a more detailed study of the residuals. We will illustrate diagnostic checking of residuals in the examples in Chapter 6.

2.5 Forecasting

We now consider the necessary results to use VARMA models for forecasting.

2.5.1 Minimum mean square error forecasts

Assuming the model form and true parameter values to be known, we can produce forecasts of the future values of our k variables at times $t+h$ made from time

origin t (denoted by $\hat{\mathbf{z}}_t(h)$ – h is called the **horizon**). These have associated forecast errors $\hat{\mathbf{e}}_t(h)=\mathbf{z}_{t+h}-\hat{\mathbf{z}}_t(h)$ which have positive definite covariance matrix $V(h)$. Granger and Newbold (1986) consider minimising the forecast error covariance matrix $V(h)$ in the sense that a square matrix A is said to be positive (definite) if $\mathbf{x}^TA\mathbf{x}>0$, $\forall\mathbf{x}\neq\mathbf{0}$; so that $V_1(h)>V_2(h)$ if $\mathbf{x}^T[V_1(h)-V_2(h)]\mathbf{x}>0$, or $\mathbf{x}^TV_1(h)\mathbf{x}>\mathbf{x}^TV_2(h)\mathbf{x}$, $\forall\mathbf{x}\neq\mathbf{0}$ and hence we can find a **minimum** forecast error variance matrix $V(h)$. The forecasts are taken to be linear sums of past known $\mathbf{z}_t$s

$$\hat{\mathbf{z}}_t(h)=\sum_{i=0}^{\infty}\lambda_{h,i}\mathbf{z}_{t-i} \qquad [2.26]$$

$$=\sum_{i=0}^{\infty}\lambda_{h,i}\sum_{j=0}^{\infty}\Psi_j\mathbf{a}_{t-i-j}, \quad \text{MA}(\infty)\text{ form}$$

$$=\sum_{k=0}^{\infty}\Lambda_{h,k}\mathbf{a}_{t-k} \qquad (\Lambda_{h,k}=\sum_{j=0}^{k}\lambda_{h,j}\Psi_{k-j})$$

(this assumes that the VARMA model for $\mathbf{z}_t$ is stationary and so the MA(∞) form exists). The forecast errors from these forecasts are then

$$\hat{\mathbf{e}}_t(h)=\mathbf{z}_{t+h}-\hat{\mathbf{z}}_t(h)$$

$$=\sum_{j=0}^{\infty}\Psi_j\mathbf{a}_{t+h-j}-\sum_{k=0}^{\infty}\Lambda_{h,k}\mathbf{a}_{t-k}$$

$$=\sum_{j=0}^{h-1}\Psi_j\mathbf{a}_{t+h-j}+\sum_{j=0}^{\infty}(\Psi_{j+h}-\Lambda_{h,j})\mathbf{a}_{t-j}.$$

This gives us

$$V(h)=\sum_{j=0}^{h-1}\Psi_j\Sigma_a\Psi_j^T+\sum_{j=0}^{\infty}(\Psi_{j+h}-\Lambda_{h,j})\Sigma_a(\Psi_{j+h}-\Lambda_{h,j})^T,$$

which has a smallest value of $V(h)=\sum_{j=0}^{h-1}\Psi_j\Sigma_a\Psi_j^T$ when $\Lambda_{h,j}=\Psi_{j+h}$, i.e. the forecasts are

$$\hat{\mathbf{z}}_t(h)=\sum_{k=0}^{\infty}\Psi_{h+k}\mathbf{a}_{t-k}. \qquad [2.27]$$

The MA(∞) form can be recursively generated from

$$\Phi(B)\mathbf{z}_t=\Theta(B)\mathbf{a}_t=\Phi(B)\Psi(B)\mathbf{a}_t,$$

by equating powers of B on either side of the last equality to give

$$\Psi_1=\Phi_1-\Theta_1, \quad \Psi_2=\Phi_1\Psi_1+\Phi_2-\Theta_2 \quad \text{etc.}$$

With our model in the form [2.1] we can obtain forecasts $\hat{\mathbf{z}}_t(h)$ by recursively replacing unknown quantities on the r.h.s. by their forecast values, e.g.

$\hat{\mathbf{z}}_t(1)=\Phi_1\mathbf{z}_t+\ldots+\Phi_p\mathbf{z}_{t+1-p}+0-\Theta_1\mathbf{a}_t-\ldots-\Theta_q\mathbf{a}_{t+1-q}$, (1–step ahead)

$\hat{\mathbf{z}}_t(2)=\Phi_1\hat{\mathbf{z}}_t(1)+\Phi_2\mathbf{z}_t+\ldots+\Phi_p\mathbf{z}_{t+2-p}-\Theta_2\mathbf{a}_t-\ldots-\Theta_q\mathbf{a}_{t+2-q}$, (2–steps)

so that we can also forecast from models which do not necessarily satisfy the stationarity conditions in §2.1.3.

2.5.2 Forecast errors

The forecast errors (assuming the model form and its parameters are known exactly) and their covariances are given by

$$\hat{\mathbf{e}}_t(h)=\sum_{i=0}^{h-1}\Psi_i\mathbf{a}_{t+h-i},\quad \mathrm{var}(\hat{\mathbf{e}}_t(h))=V(h)=\sum_{i=0}^{h-1}\Psi_i\Sigma_a\Psi_i^T. \qquad [2.28]$$

We can also obtain formulae for the covariances of the forecast errors from models in the form [2.1] recursively as shown below.

	Forecast error	covariance
$\hat{\mathbf{e}}_t(1)=$	$\mathbf{a}_{t+1}$	Σ_a
$\hat{\mathbf{e}}_t(2)=$	$\Phi_1\hat{\mathbf{e}}_t(1)+\mathbf{a}_{t+2}-\Theta_1\mathbf{a}_{t+1}$	
$=$	$(\Phi_1-\Theta_1)\mathbf{a}_{t+1}+\mathbf{a}_{t+2}$	$(\Phi_1-\Theta_1)\Sigma_a(\Phi_1-\Theta_1)^T+\Sigma_a$

We may be interested in the covariance (or correlation) between forecast errors from forecasts made at different horizons h_1 and h_2. This can be derived using

$$\hat{\mathbf{e}}_t(h_j)=\sum_{i=0}^{h_j-1}\Psi_i\mathbf{a}_{t+h_j-i},\quad j=1,2$$

and thus the covariance between the errors $\hat{\mathbf{e}}_t(h_1)$ and $\hat{\mathbf{e}}_t(h_2)$ is given by (assuming without loss of generality that $h_2>h_1$)

$$E[\hat{\mathbf{e}}_t(h_1)[\hat{\mathbf{e}}_t(h_2)]^T]=\sum_{i=0}^{h_1}\Psi_i\Sigma_a\Psi_{h_2-h_1+i}^T \qquad [2.29]$$

(since $\mathbf{a}_t$ is white noise and $E[\hat{\mathbf{e}}_t(h)]=0$ – see [2.28]), which corresponds with the variance in [2.28] when $h_1=h_2$. The correlation between forecasts at different horizons can be obtained by scaling [2.29] using the variances from [2.28].

2.5.3 Estimated models

When the actual model form is unknown and the parameters have to be estimated, further uncertainty is introduced into the forecast errors and the measurement of their uncertainty, so that it is only sensible to produce forecasts from a thoroughly tested and satisfactory VARMA model. Forecasting properties of VARMA models do not appear to have been thoroughly studied although some comparative case studies have been published (see §3.4.2). The problems afflicting univariate

forecasting also carry through to the multivariate case (see for instance Chatfield (1988a)). We will discuss some aspects of forecasting with VARMA models in §3.4, together with the problems of comparing the forecasting performance of different models in §3.3.2.

2.6 Other models and special cases

In this section we consider some alternative ways of modelling multivariate time series (and the situations in which we may want to use them), together with some special cases of the VARMA model.

2.6.1 Special cases

If k=1, our vector of observations becomes a scalar, and the VARMA model reduces to a univariate ARMA model. While all of the theory holds true if k=1, the representations used are aimed at the vector case and are unnecessarily cumbersome for univariate analysis, for which many, more useful tools have been developed. We could collect together k univariate models for a collection of series in a VARMA form with the parameter matrices constrained to be diagonal, but this is unlikely to be a realistic representation in practice, where we might expect some between–series effects (since we are modelling them jointly). Similarly, if all of the coefficient matrices are lower triangular, the model reduces to that of a transfer function (§1.3) since some of the variables become exogenous with no "feedback" from the others. It is likely to be better to specifically use transfer function techniques when this is the case, but, of course, we may not know the direction of causation beforehand and want to build VARMA models to explore it.

2.6.2 State–space representations

Another form to model our observed time series variables assumes an underlying **state vector** of m unobservable system parameters $\mathbf{y}_t$, which evolve according to a linear **system equation**

$$\mathbf{y}_{t+1}=F\mathbf{y}_t+\mathbf{a}_{t+1}, \quad \mathbf{a}_t \sim N_m(\mathbf{0},\Sigma_a). \qquad [2.30a]$$

These are thought to affect the observed variables through an **observation equation**

$$\mathbf{z}_t=H\mathbf{y}_t+\mathbf{b}_t, \quad \mathbf{b}_t \sim N_k(\mathbf{0},\Sigma_b). \qquad [2.30b]$$

This form was developed from control and systems theory (see for example Hannan and Diestler (1988) which includes much relevant theory, and the references therein). Many workers now use this state–space representation and for instance, Harvey (1989) presents a thorough account of building **structural** models where the state vector has interpretable components such as trend or

seasonal components (which may often be sensible). Hannan and Diestler show that under certain conditions, any state–space system ([2.30a and b]) has a VARMA representation and any VARMA model (satisfying the usual conditions) can be written in a state–space form. This correspondence allows, for example, estimation routines for state–space models to be exploited for VARMA models (§2.3.4) and also means that we can use whichever representation is most suitable for our particular application. The VARMA representation allows us to explore the system under study in terms of the original variables, with few prior restrictions on the structure we expect to find (Tiao and Tsay (1989 – reply to discussion)) without relying on a, perhaps unnatural, underlying "system".

2.6.3 Econometric models

Often economists, for example, will specify large, simultaneous equation models for a multivariate system (e.g. Zellner and Palm (1974)) which use many variables and parameters and are based upon econometric theory and prior beliefs. Mills (1990, Chapter 14), for example, discusses the relationships between various econometric models and VARMA representations. These systems of equations are thought to model the behaviour accurately, but require a great deal of experience and little statistical input to build. They rarely give better forecasts than more parsimonious models (except where the forecasts are "judgementally adjusted" – see e.g. Boero (1990)) and do not lend themselves to the exploration of relationships between variables, since these are usually specified beforehand.

2.6.4 Bayesian models

Litterman (1986) reviews some aspects of forecasting with Bayesian VAR models, while West and Harrison (1989) present much of the theory to build more general Bayesian models (called Dynamic Linear Models – multivariate considerations are given in Chapter 15 of the book). Where we have substantial prior knowledge about a system which can usefully be incorporated into a representation, it may certainly be worthwhile to exploit it with such models, but otherwise, that is when exploring relationships between variables, it might be better to have a more flexible structure which allows perhaps unexpected relationships to be revealed.

2.6.5 Non–linear models

Some observed phenomena may not be adequately described by the linear VARMA model – for example in hydrology many flows are not symmetric (and are said to be **time–irreversible**), which is implicit in the linear Gaussian formulation. Tong (1983) considers piecewise linear or threshold models with time–varying parameters to describe such behaviour, while Mills (1990, Part IV)

discusses the use of non–linear models in economic situations which can often be asymmetric (although there appears to have been little work done which is specific to the multivariate case). Inherently non–linear phenomena will best be modelled by taking this behaviour into account and we would hope to be able to detect such deviations from linearity by examining the residuals from a linear model, which might suggest that we consider non–linear models or perhaps transform the data.

There are other representations which might be useful in certain circumstances. For instance, if forecasting were the prime objective rather than modelling, then Harvey (1983) describes a multivariate exponential smoothing model, while Otter (1990) derives direct minimum mean square error forecasts for multivariate series using canonical correlation techniques (– perhaps even a collection of univariate models may be satisfactory). Pfeifer and Bodily (1990) consider the situation where our k series are the same variable measured at different sites in space and thus spatial features become important. They use a collection of univariate ARMA models tied together with spatial parameters (which could also be thought of as a VARMA model with restrictions on the coefficients). Granger and Weiss (1983) introduce the error–correction representation of non–stationary economic series which models changes in the variables based upon previous deviations of a particular linear combination of the variables from an **equilibrium** (see also co–integration in Chapter 4).

Chapter 3. Problems with VARMA modelling

In Chapter 2 we presented the VARMA model and the tools necessary to build such models to describe multivariate time series relationships. Although these are essentially all that we need, there are many other considerations and problems, some of which are unique to the multivariate case (e.g. co–integration) and some that we have met in the univariate case, but which are made worse by the inevitable increase in size of the problem (such as estimation). In this chapter we discuss some of these problems, deferring detailed discussion of two areas, co–integration and model specification until later chapters.

3.1 Deviations from assumptions

The derivation of the VARMA model form [2.2] given in §2.1.1 relies upon various assumptions made about the data being modelled. Further conditions were imposed on the representation in §2.1.3 in order to promote some sensible properties of the model. However, it can be the case that some of these restrictions do not hold in practice and we will have to transform our data in some way to satisfy them.

Probably the most commonly violated assumption is that of stationarity of the data. In the univariate case, stationarity is taken to be that the mean μ and variance σ^2 of the process z_t are constant and the autocovariance $\text{cov}(z_t z_{t+h})$ depends only on the lag h between observations. (This is the usual, second–order, definition of stationarity which is sufficient under the normal assumption. Box and Jenkins (1976, §2.1.2) for example define strict stationarity.) In the multivariate case a vector series $\mathbf{z}_t=(z_{1t},\ldots,z_{kt})^T$, is said to be (second order) stationary if **each** component series z_{it} is (univariate) stationary **and** $\text{cov}(z_{it},z_{j,t-h})$ is independent of t for all i,j,h (joint covariance stationary – §2.1.1).

3.1.1 Univariate considerations

There are many ways in which a **single** series can deviate from stationarity – in the mean, variance, covariance or some combination of these. Examples of some non–stationary models which are thought to approximate the behaviour of series found in practice include the following.

(a) Deterministic trend

The mean of the process may not be constant (different subsets of the data have different local means), that is $\mu=E[z_t]=f(t)=\mu_t$ or $z_t=f(t)+a_t$. The mean can be taken to follow some deterministic function of time $f(t)$, typically linear or low degree polynomial. $f(t)$ can also be a trigonometric or other periodic function

which then accounts for some seasonal or cyclic behaviour in the series. Although it is tempting to try and separate out deterministic components of a series' behaviour in this way (using, for instance, the classical trend, seasonal and stationary stochastic decomposition – see e.g. Granger and Newbold (1986) or Wei (1990)) it is unlikely that such components are entirely deterministic and the approximation may be considerably worse outside of the region of fit. We can consider a stochastic trend term in, for example, the structural model of Harvey (1989), but in (c) below, we consider differencing which can be a flexible way of dealing with non–stationarity in the mean for ARMA models.

(b) Non–constant variance

The size of the fluctuations in economic variables can commonly be related to the current level of the series in some way. A possible model for this may be

$$z_t=T(t)y_t,$$

where y_t is a stationary process and $T(t)$ a deterministic function of time. Logarithmic transformation of z_t (perhaps after adding a constant onto the values to make them all positive) will then give a series with a deterministic trend component which can be modelled using e.g. (a) above. Box and Cox (1964) considered the case where

$$\text{var}(z_t)=cf(\mu_t),$$

for some positive constant c and suitable function f, and we seek a (non–linear) transformation g such that $g(z_t)$ has constant variance. It can be shown, using a Taylor series expansion, that the function g is given by

$$g(x)=\int\frac{1}{\sqrt{f(x)}}\,\text{dx}.$$

If, for example, the standard deviation is proportional to the level i.e. $f(\mu_t)=\mu_t^2$, then a logarithmic transformation will give a series with constant variance. Generally, we choose a variance stabilising transformation from the "power" family

$$g(z_t,\lambda)=z_t^{(\lambda)}=\frac{z_t^{\lambda}-1}{\lambda}, \qquad [3.1]$$

for some parameter λ (chosen to control the variance before further analysis, or estimated as a parameter in the likelihood function of a model for the data – see e.g. Wei (1990, Chapter 4)).

(c) Homogeneous non–stationarity; differencing

Box and Jenkins (1976) suggested that often the non–stationary behaviour will be

in some way **homogeneous** through time. That is, different parts of the series will behave similarly except for their difference in, for example, local mean levels. Examples of this are shown in Figure 3.1. The series in Figure 3.1a was generated by

$$z_t=z_{t-1}+a_t,\ \ a_t\sim N(0,1) \qquad [3.2]$$

and the boxes show regions which appear to behave similarly except for their difference in level. The series in Figure 3.1b was generated by

$$z_t=2z_{t-1}-z_{t-2}+a_t,\ \ a_t\sim N(0,1) \qquad [3.3]$$

and the boxes show apparently similar regions, differing only in level and slope. If we model the series generated by [3.2] with an ARMA model, we require the model to be such that

$$\psi(B)(z_t+c)=\psi(B)z_t,\ \ \forall c$$

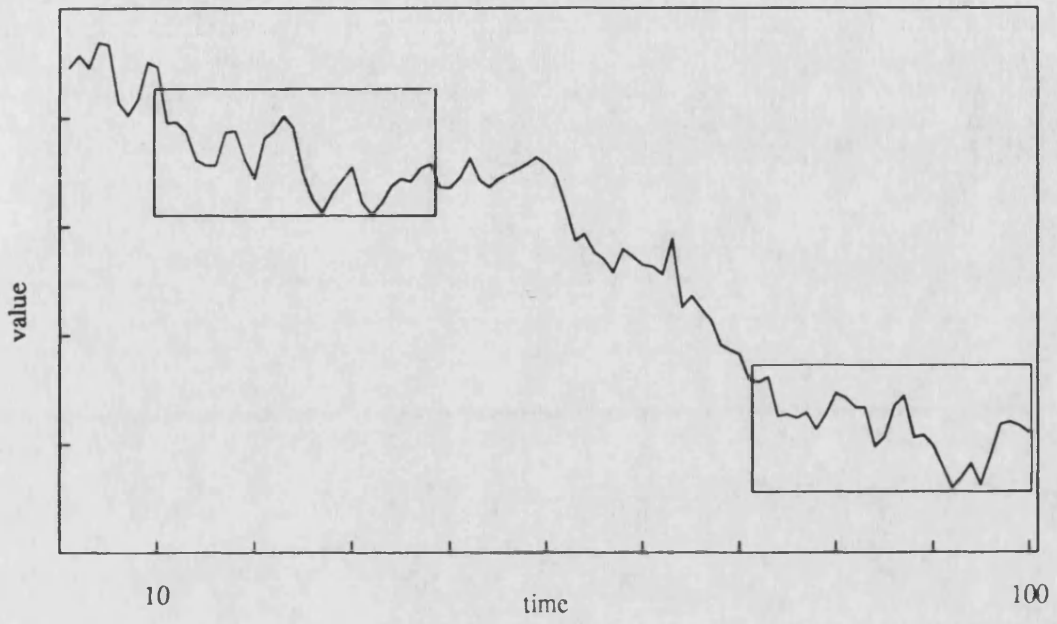
(where $\psi(B)$ is an autoregressive operator), which implies that $\psi(B)=\varphi(B)(1-B)^d$ (for some stationary $\varphi(B)$ and $d>0$), since $(1-B)^d(z_t+c)=(1-B)^d z_t$. This then leads us naturally to taking differences ($\nabla=1-B$) of series to reduce them to stationarity. A series is said to be **integrated** to degree d, or I(d), if it requires d'th degree differencing to make it stationary. We must be careful not to difference too much, or we will encounter problems due to this overdifferencing – see §3.1.2. For z_t generated by [3.2], $\nabla z_t=a_t$ is stationary, so z_t is I(1), while for z_t generated by [3.3], $\nabla^2 z_t=a_t$ is stationary – z_t is I(2).

So far, with stationary series, we have subtracted the (constant) mean level away from the observations to simplify the modelling process. With series which are non–stationary in the mean, this is not sensible, but the first degree differencing operator corrects for a stochastic local level. In the general univariate ARMA model [1.1] of §1.2.1, we can include a constant term θ_0 on the right hand side of the difference equation and consider possibly differencing z_t d times (an ARIMA model). If $d=0$ (a stationary process), then θ_0 corresponds to $(1-\varphi_1-\ldots-\varphi_p)\mu$, otherwise it is a deterministic trend or drift term (since the increases have a fixed, non–zero mean value θ_0). To see this, we can modify [3.2] to give us

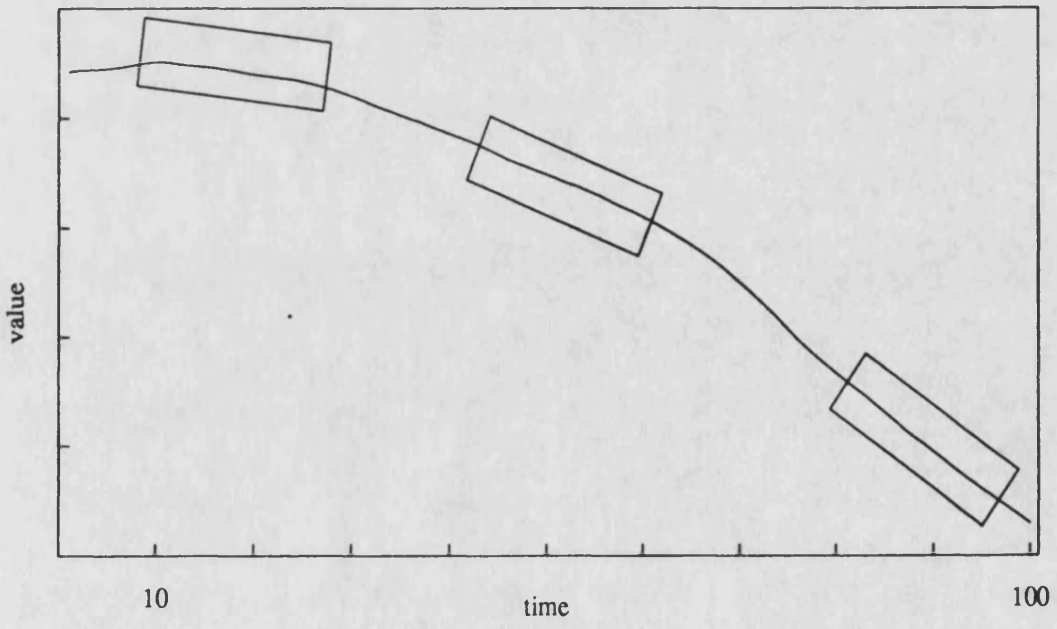
$$\begin{aligned} z_t&=z_{t-1}+\theta_0+a_t \qquad [3.4]\\ &=z_{t-2}+2\theta_0+a_t+a_{t-1}\\ &\ldots\\ &=z_{t-h}+h\theta_0+\sum_{j=0}^{h}a_{t-h},\ \ h>0, \end{aligned}$$

Figure 3.1

(a) I(1) series [3.2]



(b) I(2) series [3.3]



(applying [3.4] recursively for different times t), illustrating the deterministic nature of the θ_0 term. This behaviour will dominate for large t (see Figure 3.2 which shows two series generated by [3.4], with $\theta_0=0.1$ in 3.2a and $\theta_0=1$ in 3.2b – $\sigma_a^2=1$ in both), but, for the example in [3.4], first degree differencing results in a stationary series with non–zero mean (θ_0) which can be subtracted. With reference to (a) above, if the mean of the process follows a d'th degree polynomial trend, then differencing of degree d will produce a series without trend, but with a non–zero mean θ_0.

We will consider taking a variance stabilising transformation ([3.1]) and then differencing to reduce the observed series to stationarity. We look at differencing in a multivariate context in §3.1.2. Univariate (homogeneous) non–stationary behaviour is suspected in practice from an inspection of the sample acf (e.g. Box and Jenkins (1976, Chapter 6)), specifically, if this function does not die out quickly. There are procedures which can be applied to test the hypothesis that the series requires differencing (i.e. that the model has a unit root – see e.g. Dickey, Bell and Miller (1986)).

(d) Seasonality

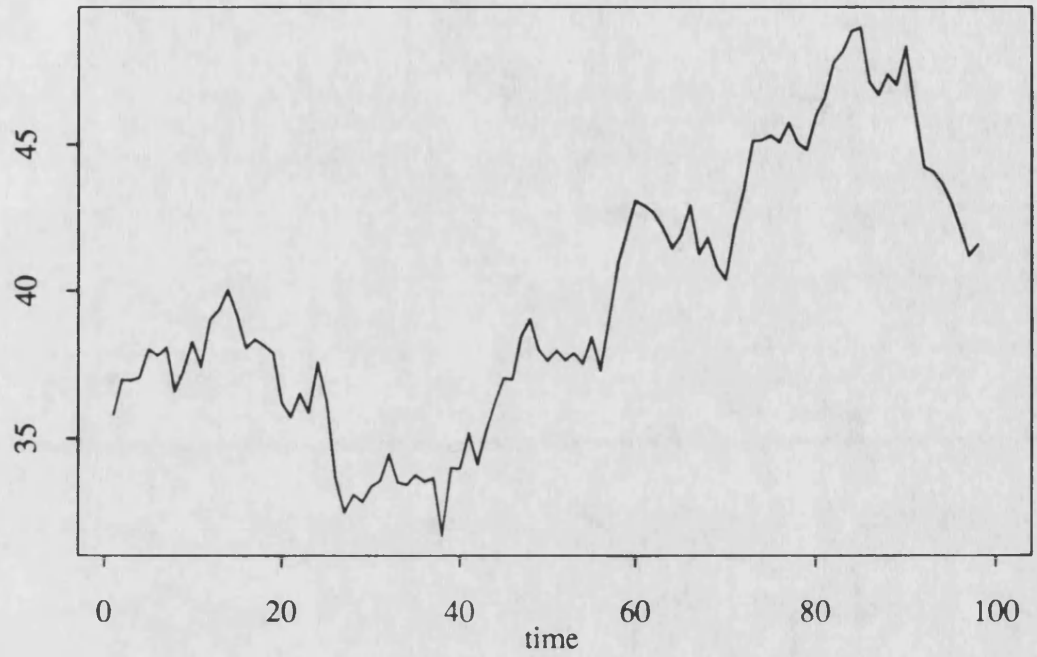
As mentioned in (a) above, we can model seasonal behaviour with a deterministic periodic component, but more generally (since the seasonal behaviour is unlikely to be entirely deterministic) we might include a seasonal term in our stochastic model for the data (see §2.1.2 and §5.2.11; also Gooijer and Klein (1989) for examples). It could happen though, that our series is such that z_t-z_{t-s} is stationary for some fixed seasonal period s (whilst z_t is not stationary) and we are led to consider seasonal differences ($\nabla_s=1-B^s$, differencing of order s to varying degrees). As with ordinary (first order) non–stationarity, the possible need for such treatment is diagnosed from inspection of the sample acf (at lags which are multiples of s). See Hillmer and Tiao (1982) or Wei (1990, Chapter 8) and the references therein for more details of the effects of seasonality. As mentioned in §2.1.2, multivariate seasonal considerations can become very cumbersome, in which case it may be sensible to use a seasonal adjustment procedure, such as are frequently used for many economic variables (§6.3, but see Jenkins (1979)).

3.1.2 Multivariate differencing

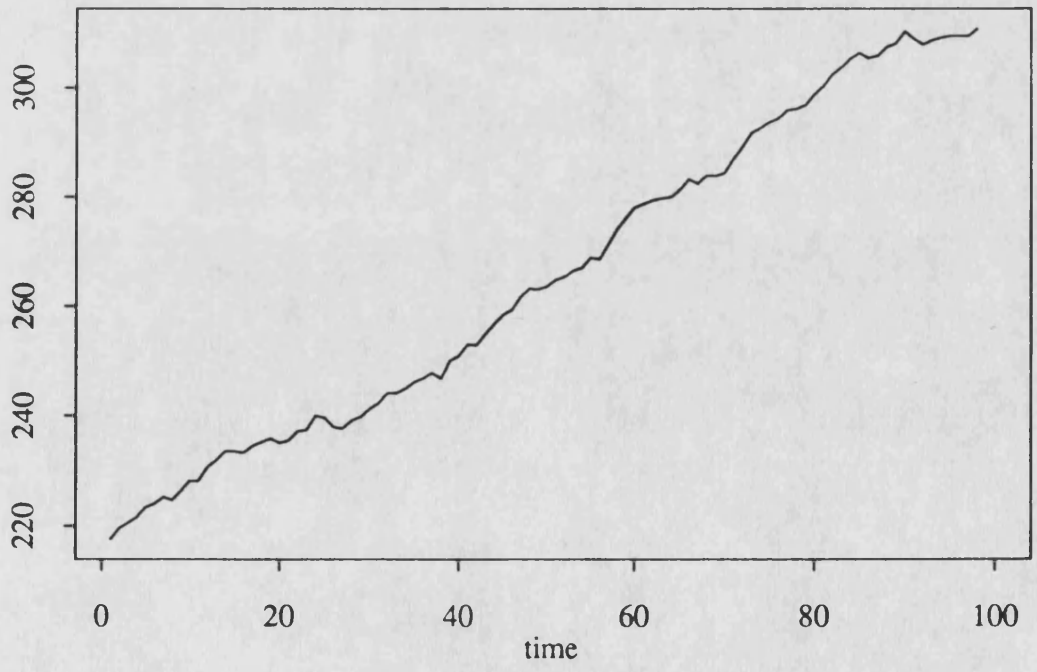
In the vector case we have $\mathbf{z}_t=(z_{1t},\ldots,z_{kt})^T$ and consider first non–linear transformations of the form [3.1] for each variable z_{it} to stabilise the variance, and then possibly differencing each variable in an effort to reduce them to stationarity. We first meet a problem since each series may not require the same

Figure 3.2

(a) series [3.4], θ_0=0.1



(b) series [3.4], θ_0=1



transformation parameter nor the same degree of differencing and so the model necessarily becomes more complicated. For instance, we may suspect that individually, series i requires a power transformation with parameter λ_i and then differencing of degree d_i. We can then analyse the (multivariate) stationary series $D(B)\mathbf{z}_t^{(\lambda)}$ where $\mathbf{z}_t^{(\lambda)}=(g(z_{1t},\lambda_1),\ldots,g(z_{kt},\lambda_k))^T$ are (power) transformed series and

$$D(B)=\begin{bmatrix}(1-B)^{d_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & (1-B)^{d_k}\end{bmatrix} \qquad [3.5]$$

is a differencing operator. Care must be taken when applying operators like [3.5] to try and reduce multivariate series to stationarity, since it may be that there exist (perhaps several) stable contemporaneous relationships between unstable variables (e.g. Engle and Granger (1987)). This is termed **co–integration** and we consider the implications of this for VARMA models, including ways of testing and correcting for it in Chapter 4. In the presence of co–integration, differencing such as [3.5] can be unnecessary even though each series individually requires differencing to degree d_i, and it will lead to overdifferencing if applied without first examining the contemporaneous relationships between series (see the example in §4.2). Models for overdifferenced data will include unnecessary unit roots in the MA polynomial which can induce a large bias in the parameter estimates (Hillmer and Tiao (1979)) and so should be avoided. Tiao and Box (1981) allow the zeroes of the determinantal polynomial $|\Phi(B)|$ (see §2.1.3) to lie on the unit circle in order to account for the existence of non–stationarity and avoid difficulties due to overdifferencing. However, although we only require $|\Phi(0)|\neq 0$ to obtain a state space representation of a VARMA process (Hannan and Diestler (1988), which is certainly the case in our representation [2.2]), the representations used in the algorithms of Ansley and Kohn (1983) and Shea (1987) require $|\Phi(z)|\neq 0$ for $|z|\leq 1$, so that Tiao and Box's suggestion is not feasible when using such algorithms for estimation. We will demonstrate the approach which we recommend for dealing with non–stationarity in multivariate series in §4.2.3 and §4.5.

3.1.3 Interventions and outliers

Frequently, external events such as strikes, policy changes etc. affect the structure generating the variables under observation and their analysis and the approximation to many of the modelling assumptions can be improved by taking these changes into account. If we know the time at which the event happened (T, say), then it is termed an **intervention**. We can assume that the effect is either in the form of a step, i.e. permanent change to a single variable

$$S_t^{(T)} = \begin{cases} 0 & t<T \\ 1 & t \geq T \end{cases}, \qquad [3.6]$$

which may be sensible for policy changes, for example, or a pulse change

$$P_t^{(T)} = \begin{cases} 1 & t=T \\ 0 & t \neq T \end{cases}. \qquad [3.7]$$

which should better describe isolated events such as strikes (although we could also consider other effects, such as a decaying step $=1/(t-T+1)$, $t \geq T$). These intervention (dummy) variables can then be included in models for the individual variables through a transfer function (§1.3). In a multivariate model, we can incorporate the effects of suspected interventions on each of the series with S_t or P_t as exogenous variables in an ARMAX model (Hannan and Diestler (1988)) or as additional variables with constraints on the parameter matrices to enforce their exogeneity (§2.1.2). Box and Tiao (1975) among others, deal with the analysis of the effects of such known interventions.

When the timing and causes of such external events is not known, but their effects are suspected from, for example, an analysis of the data or residuals from a model fitted to it, then these effects can be accounted for with **outlier** models. Fox (1972) proposed two models for outliers, both of which perturb an underlying outlier-free series to produce the observed series. Tsay (1986), Chang, Tiao and Chen (1988) and Bruce and Martin (1989) look at aspects of such outlier models. It is easy to "correct" for suspected outliers in order to improve features of a model (see e.g. Bhattacharyya (1982), where 7 interventions are fitted in a model for series of length 60, in order to improve the fit), but unless evidence can be found to support the suspicion, it can lead to overspecified models. If good evidence is found, then an intervention variable should perhaps be used to model the effects in an interpretable way. In a multivariate context, many of the outlier detection methods will become complicated to apply, nevertheless, interventions can be important in several series (see the background to the datasets of §A.1 and §A.4).

3.2 Model specification and the large number of parameters

The wide range of possible VARMA structures worsens the univariate problem of choosing a suitable model to represent observed data. Also any chosen model (order) will often have a large number of parameters which need to be estimated. In this section we discuss these two problems in the context of VARMA models – model specification procedures aimed at tackling both problems will be considered in detail in Chapter 5.

3.2.1 Model Identification

The model identification methods described in §2.2, which are direct multivariate generalisations of the univariate methods of Box and Jenkins, become difficult to apply in practice when the number of series and the complexity of structure increase. Cut–offs in the correlation matrices can be difficult to spot and mixed ARMA model identification is not straightforward, even using the tools discussed in §2.2.3. One difficulty is that of deciding when a matrix can be regarded as zero and many of the yardsticks proposed in §2.2 serve only to suggest tentative initial model orders which are worthy of further investigation. In the univariate case, Gooijer *et al.* (1985) survey many methods of order determination which supplement those usually applied, some of which may be applicable (or generalised) to the multivariate case (but see §5.1 and §5.6). In the transfer function case (exogenous inputs) outlined in §1.3.2, we used the technique of prewhitening (filtering the series by a univariate model for the input) to aid model identification. In the more general VARMA case without exogenous inputs, it is not necessarily sensible to prewhiten using a model for any particular series. Haugh and Box (1977) propose another prewhitening method which involves analysing the residuals from univariate models for **each** of the series, claiming that their ccf will be the least corrupted by autocorrelation within the individual series and hence easiest to interpret. Although appealing, any prewhitening technique will make the analysis extremely complicated and can lead to overparameterisation (Tiao and Box (1981)). This also applies to the technique mentioned in §2.4, of using a model for the residuals to modify the original model for the data (see also Box and Jenkins (1976, Chapter 8)). The combination of the models is prone to being overparameterised with possible simplifications and common factors difficult to spot and the technique is not so applicable to the multivariate as the univariate case. It is better to use any inadequacies in the residuals as guides for modifying the original model.

3.2.2 Estimation

The estimation of large multivariate models is ill–conditioned and time–consuming and can result in highly correlated and unstable estimates. To make it more manageable, good starting points are required and constraints may have to be placed upon the parameters. The nature of the VARMA model [2.2] means that we introduce k^2 extra parameters for every increment in model order that we choose to specify, so it is essential to keep the orders p and q as low as possible and also to detect and remove insignificant parameters. Various methods have been developed to help specify parsimonious **univariate** ARMA models (Piccolo and Tunnicliffe Wilson (1984) survey many of them), some of which generalise to

the multivariate case.

3.2.3 Parsimonious parameterisation

One way to reduce the size of the multivariate problem is to reduce the effective **dimension** (k) of the system under study. Several techniques from multivariate analysis have been applied to the time series (i.e. correlated) problem, including principal components analysis (Priestley *et al.* (1974) – see also Priestley (1981)) and factor analysis (Peña and Box (1987)) and these will be applicable to systems where dimension reduction is likely to be feasible. More generally, parameter redundancy occurs in the coefficient matrices Φ_i and Θ_i and good initial estimates of the parameters can help to identify and remove insignificant entries before attempting an ill–conditioned maximum likelihood estimation. Koreisha and Pukkila (1987) proposed an *ad hoc.* method for identifying insignificant entries in the coefficient matrices, but it can be more fruitful to transform the data so as to exploit possible linear dependencies among rows of the parameter matrices (see e.g. §5.1.1). There are many methods which attempt to do this in order to reduce the size of the estimation problem and they usually provide identification of a suitable model based upon the minimal specification of components of the system. We consider some of these methods in Chapter 5 and give examples of their application in Chapter 6.

3.3 Model comparison

Once we have selected and fitted suitable models for our data, we will then usually want to compare them in some way and perhaps choose between them. The use of most model specification procedures will result in a choice of model order, at least in some cases, and we may also want to compare other models which we have available (such as univariate or white noise models). The purpose for which the model was developed often suggests specific criteria to use – for instance, if forecasting were a prime objective, then it is most sensible to use out–of–fit forecast performance to compare different models. Gooijer *et al.* (1985) mention a form of cross–validatory checking which involves estimating a deleted observation, but it is probably too expensive to apply, particularly to multivariate series. Also we are usually interested in extrapolation rather than interpolation in time series contexts.

It could happen that models are in fact similar (in effects and inferences) although they appear different (in structure and parameters). For this reason, it can be useful to examine the AR(∞) and MA(∞) representations which will be unique for a given (stationary etc.) process (§2.1.1). In the univariate case, Piccolo (1990) has proposed a distance measure for comparing models and defines

$$d(X,Y)=(\sum_{j=1}^{\infty}(\pi_{j,X}-\pi_{j,Y})^2)^{\frac{1}{2}} \qquad [3.8]$$

to be the "distance" between models X and Y, where $\pi_{j,X}$ and $\pi_{j,Y}$ are the parameters from the AR(∞) representations of models X and Y respectively (model X: $(1-\pi_{1,X}B-\pi_{2,X}B^2-\ldots)z_t=a_t$). This provides an interesting tool with which to compare models using many multivariate analysis techniques and although no specific mention is made, a multivariate generalisation could use, for example $|\Pi_{j,X}-\Pi_{j,Y}|^2$ ($\Pi_{j,X}$ a parameter matrix from the VAR(∞) representation of model X), and may be worthy of further research.

3.3.1 Information criteria

Various criteria have been proposed which attempt to summarise residual information in a single value which can then be easily used for model comparison (and, as suggested by some authors, even for iterative model selection). While such a single value cannot in any way contain all of the features of a model, it can be a useful measure of fit.

Akaike (1976, for example) proposed the AIC criterion, using information theoretic arguments, which takes the form

$$\text{AIC}= -2\ln\left\{\begin{array}{l}\text{maximised}\\ \text{likelihood}\end{array}\right\} + 2\left\{\begin{array}{l}\text{number of independently}\\ \text{adjustable parameters}\end{array}\right\} \qquad [3.9]$$

and is a measure of fit penalised by the number of parameters required to achieve it. This criterion prefers the model with the **minimum** value of AIC and a difference of $2x$, say in AIC values between models is equivalent to a difference of x parameters (as some measure of the relative size of AIC values). [3.9] is already in a form applicable to both univariate and multivariate models. For univariate ARIMA models, the value of the maximised log likelihood function collapses to the sum of a constant (dependent only on the series) and $\ln\hat{\sigma}_a^2$ (where $\hat{\sigma}_a^2$ is the variance of the estimated residuals $\hat{a}_t$ from the maximum likelihood fit). For VARMA models, the **conditional** log likelihood function has the form given in [2.24]. The term $\sum_{t=1}^{N}\mathbf{a}_t^T\Sigma_a^{-1}\mathbf{a}_t$ of this is shown by Tunnicliffe Wilson (1973) to be equal to $N\text{tr}(\tilde{\Sigma}_a\dot{\Sigma}_a^{-1})$, where $\tilde{\Sigma}_a$ and $\dot{\Sigma}_a$ are estimates of the residual covariance matrix at successive iterations of the maximisation routine. Since, at convergence, these two estimates will be equal, the log of the **maximised conditional** likelihood function for a VARMA model has the form

$$\frac{-N}{2}\left\{k(1+\ln(2\pi))+\ln|\hat{\Sigma}_a|\right\} \qquad [3.10]$$

(where $\hat{\Sigma}_a$ is the final estimate of Σ_a) and hence

$$\text{AIC}=N\ln|\hat{\Sigma}_a|+2\left\{\begin{array}{c}\text{number of}\\ \text{parameters}\end{array}\right\} \qquad [3.11]$$

(up to a constant term). Akaike (1976) suggests that the criterion [3.11] can also be used for pure VAR models, while Hannan and Diestler (1988) use only this definition as a means to determine a suitable quantity called the **McMillan degree** (see §5.3.1 or Hannan and Diestler (1988, p51)) which measures the overall order of the system. Gooijer *et al.* (1985) show that [3.11] and [3.9] are asymptotically equivalent under the normality assumption and that [3.11] may be an adequate approximation to [3.9] for large numbers of observations, N. The estimation routine of Shea (1987) implemented in NAG routine g13dcf provides the log of the maximised likelihood function directly (from the likelihood of the equivalent state space representation – see also §3.5.1), so that the form [3.9] can be easily used to compare models fitted with this routine.

There have been criticisms levelled at AIC (see e.g. Shibata (1976) or Newbold (1988)), particularly when it is used as a direct order estimation criterion, and several modifications and alternative criteria have been proposed to overcome its shortcomings. Priestley (1981) surveys some of these, but there is, as yet, no recommendation for which criterion to use in general. Many of the criticisms derive from their use in model selection, where various order models are fitted to the data (usually automatically over a grid of (p,q) values, including orders which could have been discounted due to other considerations) and the criterion values tabulated. The optimum (usually minimum) value then selects the most suitable model order using that criterion. However, VARMA model fitting is too time–consuming to apply full maximum likelihood estimation to large numbers of (possibly inappropriate) models and we prefer the more direct methods of model specification, surveyed in Chapter 5. We will use the AIC information criterion (defined in [3.9]) in a final stage of comparison of the within–series fit of competing models (see the examples in Chapter 6).

3.3.2 Forecasting performance

If a goal of the modelling procedure is to produce forecasts of the series, then an assessment of the forecasting performance of different models (in similar settings to which the models will be employed – e.g. perhaps short or long–range forecasting) provides a measure with which to compare them. This is particularly of interest when we wish to see if complex multivariate models can produce more accurate forecasts than simpler, disconnected univariate ones (see §3.4.2). The comparison of univariate forecasts is by no means straightforward (see, for

example, the literature surrounding the M–competition – Makridakis *et al.* (1984) and Chatfield (1988b and c)), however most of the literature is concerned with comparing various forecasting methods over a range of different series. In this section, we are only concerned with one realisation of a collection of series and several slightly different models, perhaps produced by similar methods.

The usual approach to forecast comparison is to build and fit a model to the first N available observations and use this to forecast the remaining F (where we have $N+F$ observations on each of the k series. F is determined by the application, but could typically be about one tenth of the total available observations). We then have forecasts and their associated (known) forecast errors for each of the models, e.g. $\hat{\mathbf{z}}_N^{[A]}(h)$ and $\hat{\mathbf{e}}_N^{[A]}(h)=\mathbf{z}_{N+h}-\hat{\mathbf{z}}_N^{[A]}(h)$, $h=1,\ldots,F$, for model A, say and want to compare $\hat{\mathbf{e}}_N^{[A]}(h)$ with $\hat{\mathbf{e}}_N^{[B]}(h)$. A plot of the forecast errors from each model for each of the series can be useful for comparison and also to spot systematic behaviour which may indicate inadequacies in the models. We can also obtain the theoretical variances of these forecast errors, assuming the model to be true, as illustrated in §2.5.2 for known models. These are a measure of forecast precision, although we require our forecasts to be accurate as well as precise.

As an example of forecast comparison, Figure 3.3 shows the results for two models, A and B (models [7] and [8] of §6.2) for the Flour price data (see §A.2 and §6.2). The forecasts shown are of the last 11 observations on the first component only. Figure 3.3a shows the actual forecast errors, 3.3b shows the theoretical standard errors of these forecast errors, while 3.3c and d show the actual data and the forecasts together with their ±2× standard error limits from both models. Plots such as these provide useful indications of forecasting performance, however we would often like a way of summarising the information from such plots. A possible criterion with which to measure the actual forecast errors is "Mean Square Error", defined in the **univariate** case to be

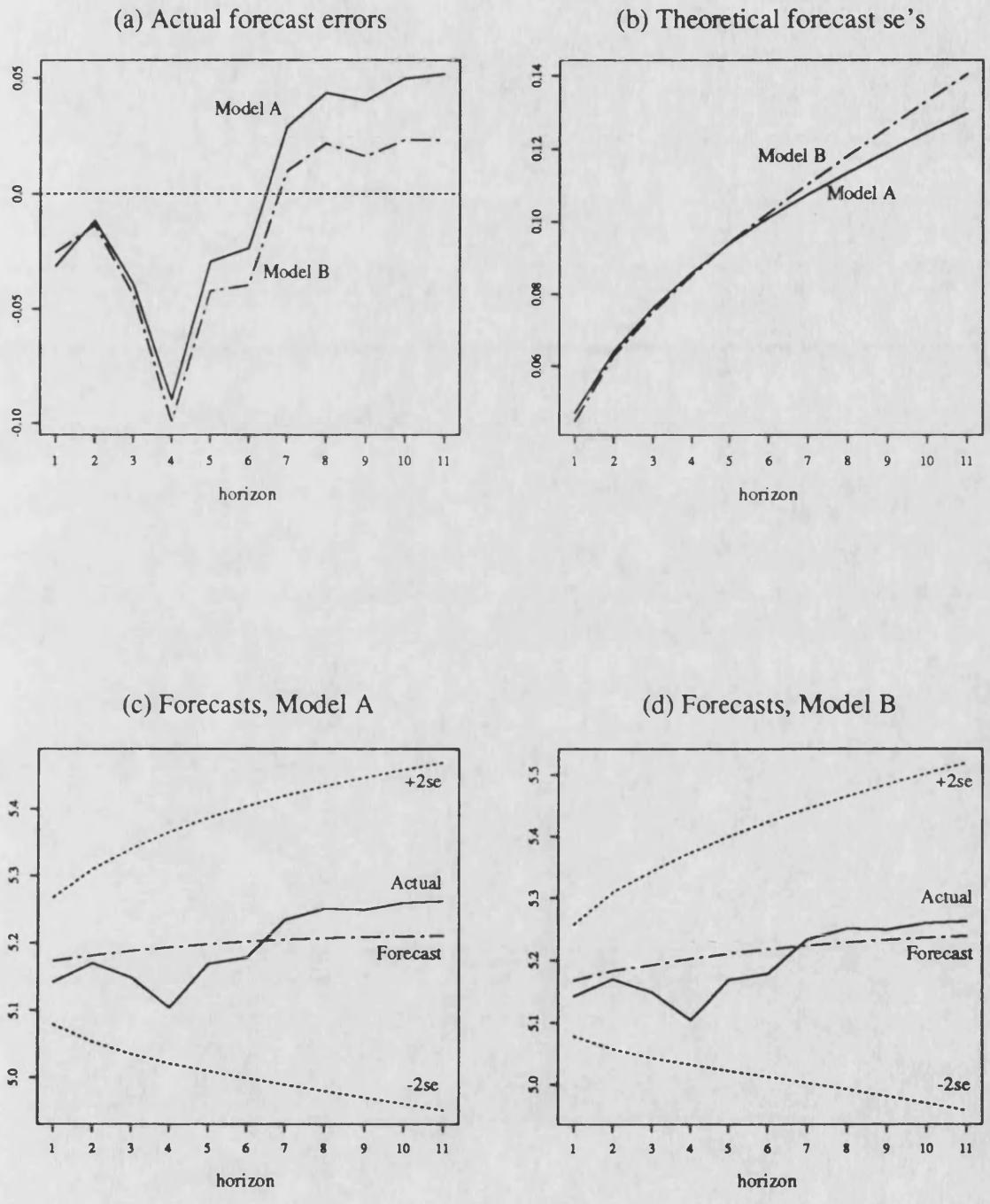
$$\mathrm{MSE}(F)=\sum_{h=1}^{F}\left[\hat{e}_N(h)\right]^2/F, \qquad [3.12]$$

so that we prefer the model which minimises this criterion for the horizon(s), F, of interest. In the multivariate case we can then consider

$$MSE(F)=\sum_{i=1}^{k}\sum_{h=1}^{F}[\hat{e}_{iN}(h)]^2/kF \qquad [3.13]$$

(the sum of the univariate MSEs over the k components), which collapses to [3.12] in the univariate case. This is only a sensible thing to consider when the component series are all comparable in scale and variability. Also the forecast errors are correlated across series – see §2.5.2. MSE has been criticised since it is

Figure 3.3


(a) Actual forecast errors
Model A
Model B
horizon
(b) Theoretical forecast se's
Model B
Model A
horizon
(c) Forecasts, Model A
+2se
Actual
Forecast
-2se
horizon
(d) Forecasts, Model B
+2se
Actual
Forecast
-2se
horizon

not scale invariant (e.g. Chatfield (1988c)) and other criteria have been suggested to try and overcome this inadequacy (Granger and Newbold (1986, Chapter 9) consider the evaluation of forecasts, while Fildes (1989) compares many criteria). These include mean absolute error (MAE) which does not penalise large errors so much as MSE, mean absolute percentage error (MAPE) which gives a scale–free measure (see e.g. Wei (1990) for definitions of these) and Thiel's U statistic (Thiel (1966)) which provides a comparison of the forecast errors with those from a "no change" (i.e. constant) forecast. A recent, appealing suggestion is Geometric Mean Square Error defined in the univariate case to be

$$\text{GMSE}=\left[\prod_{h=1}^{F}\left[\hat{e}_N(h)\right]^2\right]^{1/F} \qquad \cdots\cdots \quad [3.14]$$

(Fildes (1989)), which retains the squared penalty function of MSE, but is scale invariant and easier to interpret and compare between models. More work is needed on such criteria (see also §3.4.2, and §3.5.2 where we consider scaling to alleviate many problems), but we will generally use MSE as defined in [3.13], together with plots such as Figure 3.3, when comparing the forecasting performance of competing models (See Chapter 6 for examples).

To summarise the results about the theoretical precision of the forecasts (Figure 3.3b) we could use e.g. $|V(h)|$ or $\text{tr}V(h)$ (where $V(h)$ is the theoretical covariance matrix of the forecast errors), although it may be preferable to use plots such as Figures 3.3b and c in order to see more general features of the forecast error variances.

3.4 Forecasting

In §2.5, we showed how a VARMA model can be extrapolated into the future. This is a straightforward procedure so that (properly specified) multivariate models can easily be used to produce forecasts. We now consider some aspects of forecasting with VARMA models.

3.4.1 Exogenous variables

Often, in econometric situations, we may have exogenous variables $\mathbf{x}_t$, which can be forecast accurately outside of the system of interest, $\mathbf{z}_t$, or the modelled effect is such that we have a useful lead–time, h, between $\mathbf{x}_t$ affecting $\mathbf{z}_{t+h}$ (typically only of a few time intervals), in which case $\mathbf{x}_t$ are termed leading indicators. These effects can be incorporated into the VARMA model as suggested in §2.1.2 and the forecasts produced, conditional both on past $\mathbf{z}_t$ and past and known "future" $\mathbf{x}_t$ (termed *ex post* forecasting – Chatfield (1989)). However, we are usually interested in exploring the (unknown) relationships between the variables

under examination and may not be able to specify exogenous variables beforehand. Also, in the VARMA model, exogenous variables can be considered jointly with the other variables and so any forecasts made will often be conditional only on the past of both $\mathbf{z}_t$ and $\mathbf{x}_t$ (termed *ex ante* forecasting). Ashley (1983) considers this in the univariate case and concludes that if the exogenous variable cannot itself be forecast sufficiently accurately (defined in terms of MSE – [3.12]), then the inclusion of it in a model for (univariate) z_t will worsen the forecasts of z_t (i.e. increase their MSE). The converse of this does not hold, except when x_t is known (i.e. it is a leading indicator), in which case, its inclusion cannot worsen the forecasts of z_t. If x_t is optimally forecast (i.e. minimises the mean square forecast error, as it hopefully will, although, poor judgemental or even econometric forecasts may sometimes not be optimal) then the condition for worsening $\hat{z}_t$ will not hold and we cannot say what the effect of including $\hat{x}_t$ will be. These results confirm the considerations of §1.3.3 for transfer function models (which we would usually build when we have exogenous variables), but are not immediately useful for the more general VARMA case.

3.4.2 Forecast comparison

As soon as we start to build multivariate models, it becomes interesting to compare the forecasts with those produced by various other methods. In §3.3.2 we considered this as a way of comparing different VARMA models for the same data, but of more general importance are investigations to determine when and why forecasts from multivariate models can be better than, for instance, univariate ones. As mentioned in §3.3.2, forecast comparison is not yet a resolved issue in the univariate case, and it becomes a much more difficult problem with more complex models. Chatfield (1988b) discusses the forecasting aspects of multivariate models, whilst Ord (1988) provides a tentative summary of recent findings. There have been many studies published on the relative performance of VARMA and other forecasting methods and we summarise some recent results here.

Boero (1990) compares the *ex ante* forecasts from a structural econometric model (SEM) of the Italian economy with those from VAR models for 6 of the 600 macroeconomic variables used in the SEM. A Bayesian method is used to specify the VAR parameters in an attempt to find a parsimonious VAR(4) parameterisation and both models are used to forecast the last 20 of 60 quarterly, seasonally adjusted observations. Examination of the forecast errors reveals that the VAR model provides better short–term forecasts, whilst the SEM is better able to handle departures from the modelling assumptions which occur in the longer term, since its forecasts can be judgementally adjusted. Granger and Newbold

(1986) believe that such adjustment will often be desirable to account for unquantifiable factors, but Chatfield (1988b) points out that this will be difficult to incorporate into a statistical model.

Gooijer and Klein (1989) studied 3 macroeconomic variables concerning monthly steel traffic in the port of Antwerp. Univariate ARIMA and VARMA models were built for this data and it was found that the VARMA model selected produced the most accurate forecasts of the final 12 months (relatively short term) of available data (measured by MAPE and also by a cumulative measure of accuracy defined in the paper). It was also concluded that the initial Box–Cox transform and differencing, applied to try and make the series stationary, greatly improved the multivariate forecasts.

Lin (1989) built various time series models in order to forecast hospital patient movements and concluded that in most cases, the univariate methods (Holt–Winters – see e.g. Chatfield (1978) – and Box–Jenkins) provided better forecasts than VARMA models (measured by MSE, MAE and MAPE, for 24 monthly periods on this microeconomic data). However, only the multivariate models allowed the interactions in the system to be studied, which was another aim of the analysis presented.

Heuts and Bronckers (1988) used 5 macroeconomic variables to attempt to describe the Dutch heavy truck market. VARMA models provided a better fit for all of the series (measured in residual variances) as would be expected and also produced slightly better forecasts for most of the variables (over a short lead time, measured by MAPE).

Generally, results are encouraging, in that it can be worthwhile to build multivariate models for short–term forecasting, although the combination of forecast errors for each component can quickly make the *ex ante* forecasts stray in the longer term. This can also be due to changes in the dynamic relationships and deviations from assumptions (Boero (1990)), so that the models may need more robust specification or frequent updating. It is hoped that better model specification will improve the forecasting accuracy of multivariate models. Granger and Newbold (1986, Chapter 9) argue that although some studies have shown univariate methods outperforming the forecasts from complex econometric models, the description and interpretation of the system inherent in the larger models (including VARMA models) makes them preferable in the situations for which they have been developed.

3.4.3 Aggregation

Temporal aggregation of time series data frequently occurs, where the observations are an accumulation through time of some variable. Wei (1990, Chapter 16) considers the effects of this in the univariate case, whilst Lütkepohl (1987) considers it in a multivariate context. A second form of aggregation, frequently occurring in economic analyses uses contemporal linear aggregates of some underlying variables (maybe to reduce the dimension of the problem) and it is of interest to know whether it would be better to retain the disaggregated variables. Tiao and Guttman (1980) and Kohn (1982) have considered this and discuss the "efficiency" of aggregation. Lütkepohl (1984) presents conditions under which forecasts obtained under aggregation, disaggregation and from univariate models will be equivalent. Lütkepohl (1987) concludes that to obtain forecasts of aggregated series

(a) for **known** processes, it is best to forecast at the lowest level of aggregation (since this extends the "information base" of the forecasts) and then aggregate these forecasts.

(b) if we have to estimate the process then there can be gains in accuracy from forecasting the (contemporaneous) aggregate (since aggregating forecasts combines forecast errors). However, if we can specify good, parsimonious models then we are practically in situation (a) and use of the disaggregated series can be worthwhile.

(c) aggregating univariate forecasts is optimal if the components are uncorrelated (since we then make no gain in modelling them jointly).

(d) for temporal aggregation, the choice is more dependent upon the situation, but it is generally better to use disaggregation for stock variables (i.e. measured quantities such as price) and always better for flow variables (i.e. accumulated quantities such as production).

3.5 Data scaling and transformation

We now consider the effects which transformation or scaling the variables has upon VARMA models.

3.5.1 Linear transformations

Frequently we are required to work with some linear transformation (combination) of our multivariate data. This may be either as a result of aggregation of variables (§3.4.3) or from the application of a model specification procedure or co–integration analysis (see Chapters 4 and 5). In this section we consider the

effects such transformations have on the models.

Suppose we have two representations of a multivariate process

$$\mathbf{z}_t=\Phi_1\mathbf{z}_{t-1}+\ldots+\Phi_p\mathbf{z}_{t-p}-\Theta_1\mathbf{a}_{t-1}-\ldots-\Theta_q\mathbf{a}_{t-q}+\mathbf{a}_t, \qquad [3.15]$$

where $\mathbf{a}_t \sim N_k(\mathbf{0},\Sigma_a)$ (i.e. as in [2.2]) and

$$\mathbf{y}_t=\Phi_1'\mathbf{y}_{t-1}+\ldots+\Phi_p'\mathbf{y}_{t-p}-\Theta_1'\mathbf{b}_{t-1}-\ldots-\Theta_q'\mathbf{b}_{t-q}+\mathbf{b}_t, \qquad [3.16]$$

where $\mathbf{y}_t=T\mathbf{z}_t$, $\mathbf{b}_t=T\mathbf{a}_t \sim N_k(\mathbf{0},\Sigma_b)$, $\Sigma_b=T\Sigma_a T^T$, $\Phi_i'=T\Phi_i T^{-1}$, $\Theta_i'=T\Theta_i T^{-1}$. T is a linear transformation which is assumed to be non–singular and in some circumstances may satisfy further conditions (the case of singular T is not often met in practice and does not give rise to any results useful in this section).

(a) Likelihood

Often we can estimate a model for the transformed series, [3.16] and want to transform it back to a model for the original data [3.15]. Since T is non–singular, this is the same as transforming a model for the original data into one for a transformation of the data, so for simplicity we will look at the problem from this direction. We can use equations [3.15] and [3.16] to transform models straightforwardly, but if we wish to use, for instance, AIC ([3.9]) to compare competing models, then we will need the likelihood of the transformed model. Rather than re–estimate the model in terms of the transformed data, we can calculate the likelihood of such a model as follows. Tunnicliffe Wilson (1973) gives the conditional log likelihood for a VARMA model [3.15], as

$$L= -\tfrac{1}{2}\left[K_0+N\ln|\hat{\Sigma}_a|+\sum_{t=1}^{N}\hat{\mathbf{a}}_t^T\hat{\Sigma}_a^{-1}\hat{\mathbf{a}}_t\right], \qquad [3.17]$$

where K_0 is a constant (dependent only on k, the number of series and N the number of observations on each series), $\hat{\mathbf{a}}_t$ are the estimates of the white noise series and $\hat{\Sigma}_a$ their covariance matrix. To find the log likelihood of the transformed model [3.16],

$$L_T= -\tfrac{1}{2}\left[K_0+N\ln|\hat{\Sigma}_b|+\sum_{t=1}^{N}\mathbf{b}_t^T\hat{\Sigma}_b^{-1}\hat{\mathbf{b}}_t\right], \qquad [3.18]$$

we need to rearrange [3.18] in terms of [3.17]. Now $\mathbf{b}_t\hat{\Sigma}_b^{-1}\hat{\mathbf{b}}_t= \hat{\mathbf{a}}_t^T T^T(T^T)^{-1}\hat{\Sigma}_a^{-1}T^{-1}T\hat{\mathbf{a}}_t= \hat{\mathbf{a}}_t^T\hat{\Sigma}_a^{-1}\hat{\mathbf{a}}_t$ and $|\hat{\Sigma}_b|= |T\hat{\Sigma}_a T^T|= |TT^T||\hat{\Sigma}_a|$, i.e. $\ln|\hat{\Sigma}_b|= \ln|\hat{\Sigma}_a|+\ln|TT^T|$. This then gives us

$$L_T = L-\tfrac{1}{2}N\ln|TT^T| = L-N\ln|T| . \qquad [3.19]$$

With the state space form of the model (Shea (1989)), the log likelihood takes a similar form to [3.17], with the $\hat{\mathbf{a}}_t$s being the differences between the observations

and their conditional expectations throughout the Kalman filtering algorithm, which results in the same modification, [3.19] for a transformed model. Akaike's Information Criterion as defined in [3.9] can then be adjusted by adding $2N\ln|T|$ onto the value obtained for the original parameterisation.

(b) Number of independently adjustable parameters

In the definition of AIC in [3.9], the penalty term depends upon the number of independently adjustable parameters used in the model. With a linear transformation (particularly as a reverse transformation from a parsimonious model specified by some procedure) we may have several small/redundant parameters in the matrices Φ_i' and Θ_i' of [3.16]. However, since we have not estimated them, it would not be sensible to delete them from our model, but we may be penalising the model too heavily if we count them as model parameters. It is possible to obtain estimates of the standard errors of the parameters in [3.16] from the covariance matrix of the estimates from [3.15], however this only provides a guide to the significance of the parameters. In our model comparisons (using AIC), we have chosen to count as "independently adjustable" for [3.16] only the parameters actually estimated in the representation [3.15] since the transformation T will be fixed during the (for instance, maximum likelihood) estimation of these parameters.

(c) Forecasting

Having transformed our model to [3.16] we may then be interested in forecasting from this. There are two ways of doing this:

A obtain forecasts $\hat{\mathbf{z}}_t(h)$ from [3.15] and transform them into forecasts of $\mathbf{y}_t$, $\hat{\mathbf{y}}_t^{[A]}(h)=T\hat{\mathbf{z}}_t(h)$, and

B obtain forecasts $\hat{\mathbf{y}}_t^{[B]}(h)$ directly from [3.16].

Lütkepohl (1984) claims that it always optimal to forecast future $\mathbf{y}_t$ using method A, but gives conditions for the equality of the forecasts produced by the two methods:

$$\hat{\mathbf{y}}_t^{[A]}(h) = \hat{\mathbf{y}}_t^{[B]}(h) \text{ if and only if } \Psi'(B)T = T\Psi(B) , \qquad [3.20]$$

(Theorem 2 of Lütkepohl (1984)) where $\Psi(B)$ and $\Psi'(B)$ are the MA(∞) representations of [3.15] and [3.16] respectively. Provided that all of the matrices involved are invertible (i.e. that $|\Phi'(z)|\neq 0$, $|z|<1$, which will hold in the cases we consider) then

$$\begin{aligned}\Psi'(B)&=\Phi'(B)^{-1}\Theta'(B)\\ &=(T\Phi(B)^{-1}T^{-1})(T\Theta(B)T^{-1})\end{aligned}$$

$$\Psi'(B)=T\Psi(B)T^{-1}\ ,$$

$$\text{i.e. } \Psi(B)'T=T\Psi(B)$$

and the forecasts we obtain by methods A and B in this case will be identical.

The forecast errors from a transformed model can be calculated from those of the original model

$$\hat{\mathbf{e}}_t(h)=\mathbf{z}_t-\hat{\mathbf{z}}_t(h);\ \hat{\mathbf{f}}_t(h)=\mathbf{y}_t-\hat{\mathbf{y}}_t(h)=T\mathbf{z}_t-T\hat{\mathbf{z}}_t(h)=T\hat{\mathbf{e}}_t(h), \qquad [3.21]$$

as can their variance

$$V_e(h)=E[\hat{\mathbf{e}}_t(h)[\hat{\mathbf{e}}_t(h)]^T];\ V_f(h)=TV_e(h)T^T. \qquad [3.22]$$

3.5.2 Scaling

In §3.1.1(b) non–linear transformations were used to control deviations from stationarity of the time series variables. We may also consider scaling the series in other ways.

(a) Normalisation

Given the shortcomings of the Mean Square Error criterion mentioned in §3.3.2 it may often be sensible to use the normalised (zero mean, unit variance) data for model building and then transform this back at the final stage of inference. This would also avoid many other problems, such as with the estimation routines which may have a narrow working range (e.g. the package MTS (Reilly (1987)) checks that the data is within its range). Normalising involves subtracting the mean μ from the **stationary** series and dividing each by their standard error, i.e. we work with

$$\mathbf{y}_t=S(\dot{\mathbf{z}}_t-\mu)=S\mathbf{z}_t,\ S=\text{diag}(1/\sqrt{[\Gamma(0)]_{ii}}),$$

where $\dot{\mathbf{z}}_t$ is the original data (before mean correction) and $\Gamma(0)$ is the covariance matrix of $\mathbf{z}_t$. We then have the models given in [3.15] and [3.16] with $T=S$ and all of the considerations in §3.5.1 carry through to this scaling case (or any other scaling of the variables). For instance, the log likelihood, L_T of [3.16] is given by

$$L_T=L+\frac{N}{2}\sum_{i=1}^{k}\ln([\Gamma(0)]_{ii})\ ,$$

where L is the log likelihood of [3.15].

(b) Categorical variables

We may be interested in trying to analyse time series variables measured on a categorical scale. These often arise in psychology and the social sciences, but

could also come from more common sources. In order to apply multivariate time series techniques to such data, they will need to be scaled in some way. Buuren (1990) details much of the background to this "optimal scaling" problem and describes how it can be incorporated into the usual optimisation for many multivariate techniques. For example, if we wish to fit a univariate AR(1) model to some categorical time series data x_t, the usual (least squares) approach is to choose the parameter φ_1 (see §1.2.1) to minimise $\sum_t (x_t - \varphi_1 x_{t-1})^2$. However, we now have "category loadings" for each of the categories in which the observations fall (collected into a vector $\mathbf{y}$). The optimal loadings $\hat{\mathbf{y}}$, together with the parameter $\hat{\varphi}_1$, can be jointly estimated by choosing them to minimise $\sum_t (\mathbf{i}_t^T\mathbf{y} - \varphi_1 \mathbf{i}_{t-1}^T\mathbf{y})^2$ where $\mathbf{i}_t$ is a vector with a 1 in the position corresponding to the category in which x_t lies and zeroes elsewhere. This approach also applies to many useful multivariate analysis techniques and some implementations are described in Buuren (1990) (including Box and Tiao's (1977) predictable components canonical correlation analysis – see §5.6.1). ARMA (and therefore VARMA) modelling cannot yet be handled by the approach and the convergence of the suggested Alternating Least Squares estimates is sometimes doubtful, but it is still at an early stage of development.

One example from Buuren's book applies the technique to a real valued variable (Box and Jenkins' series D which consists of readings from a chemical process). The 310 values happen to fall into 26 discrete ordered categories, with most of the readings from the middle or right of the range. The resulting optimal scaling applied is a monotonic increasing transformation which maximises the lag–1 autocorrelation and brings the extreme values closer to the centre of the range (while preserving ordering). This appears to be a useful tool to linearise data and control outliers.

3.6 Problems considered

In this chapter we have looked at some problems encountered when trying to build and use VARMA models for real data. Solutions to many of these have been discussed, mainly as generalisations of univariate cases and while there are certainly others (particularly concerned with forecasting), these are considered to be the most relevant to the current discussion. Two particular problems have been mentioned – co–integration and model specification – which will be dealt with in more detail in the following chapters.

Chapter 4. Co–integration

Co–integration has been acknowledged as a feature of many observed multivariate time series since it was first considered by Granger (1981). It amounts to the existence of a common trend among non–stationary variables (e.g. Harvey (1989)) or more generally of relationships which tie the series together. Thus, co–integrated series remain related indefinitely, despite being inherently "unstable" (as has been observed for many econometric variables). In this chapter we first define co–integration in terms of the notion of "integration" of §3.1.1(c) and then consider the effects which co–integration and its subsequent correction can have on the (VARMA) models which we may build for the data. We discuss and extend Engle and Granger's (1987) two–stage procedure for determining the nature of the relationship(s) and illustrate this with some simulated models and a case study using real data. Further considerations, possible extensions to the methods and other proposed approaches are also discussed in later sections.

4.1 Introduction and definition

Frequently time series data will be non–stationary (=n.s.) and if they are thought to be homogeneously n.s., we may try to correct this by taking differences of each series (see §3.1.1). In the multivariate case, however, as mentioned in §3.1.2, there may exist stable contemporaneous relationships between these "unstable" variables – i.e. a non–zero linear combination of the n.s. data $\mathbf{z}_t$, may be stationary. If z_{it} are each I(1) (see §3.1.1(c)) and we can find a vector α such that $x_t=\alpha^T\mathbf{z}_t$ is I(0), then $\mathbf{z}_t$ are said to be **co–integrated** or CI (an abbreviation which we use for the noun/verb/adjective etc.). More generally the series are said to be CI(d,b) if they are individually I(d), but $\exists\ \alpha$ such that the transformed series $\alpha^T\mathbf{z}_t$ is I($d-b$). The most common case will be CI(1,1) and we will consider this for most of the discussion (using the abbreviation CI when there will be no confusion) although we consider CI among series requiring higher degree or seasonal differencing to reach stationarity in §4.6.

Notice that we only consider situations where **each** z_{it} is I(1), since if e.g. z_{jt} is I(0), then we can trivially find the stable "relationship" given by $\alpha_j=(0,\ldots,\underset{(j)}{1},\ldots,0)^T$ which picks out z_{jt}. We will usually exclude from consideration in the **CI analysis**, any series which are already I(0).

A complication is introduced because there may exist r linearly independent CI vectors, α_i (which we can collect together into a $k\times r$ matrix $\alpha=[\alpha_1,\ldots,\alpha_r]$, where k is the number of series) and it is important to determine this **CI rank**, r. From the point of view of the differencing operator [3.5], r CI relationships define a

linear transformation T, such that only $k-r$ components of $T\mathbf{z}_t$ require first degree differencing (instead of the k of the original $\mathbf{z}_t$) – see §4.2.3.

4.1.1 Econometric considerations

The econometric interpretation of CI is that the linear combination x_t is a stable **equilibrium** between the variables. For instance two variables (such as income and expenditure), may be constrained to move together, perhaps by economic forces. The existence and search for such equilibria forms an important part of econometrics where they define many proposed relationships (such as between consumption and income, which incidentally are found to be CI by Engle and Granger (1987)). We may wish to explicitly model these equilibria using, for example the **error–correction** representation of Granger and Weiss (1983). Much recent work has been done on the problem of CI in a collection of variables, particularly from an economic viewpoint, since many economic variables are n.s. but thought to obey equilibria. There is still a great deal of current interest in this topic (see e.g. the special issues of the Oxford Bulletin of Economics and Statistics (1986) and the Journal of Economic Dynamics and Control (1988)) and many problems remain to be resolved.

The testing for and identification of equilibrium relationships may sometimes be sufficient, but in this work we are also concerned with the effects that CI will have on VARMA models, and the correction of these effects. Engle and Granger (1987) (=E&G) have developed a great deal of theory concerning CI, which we will present and extend here into the VARMA context. Much of their work only considered the case when k=2 and although the generalisations to higher k are straightforward, some extra problems are encountered. Other approaches have been proposed in the literature and we discuss some of these in §4.7.

4.2 Effects of CI on VARMA models

In this section we consider how the method of reducing CI series to stationarity affects the VARMA models which we can build for the data. CI can easily be incorporated into for instance the structural model as a common trend term (Harvey (1989)) or the error–correction representation of E&G which naturally arises from CI considerations, however we must impose constraints on the variables in a VARMA model in order to account for CI. To study the effects CI has on a system of variables we will need a system which is known to behave in this way. While certain collections of real variables can be shown to be CI, they will only approximately follow any models which we choose to build for them, and so we will illustrate much of the discussion with data simulated from a suitable model. Box and Tiao (1977) give an example of a co–integrated system

$$z_{1t}=z_{1,t-1}+a_{1t} \qquad [4.1a]$$

$$z_{2t}=\beta z_{1t}+a_{2t}, \quad \mathbf{a}_t \sim N_2(\mathbf{0},\Sigma_a) \qquad [4.1b]$$

so that z_{1t} and z_{2t} are each non–stationary, but $z_{2t}-\beta z_{1t}=a_{2t}$ is stationary. This system forms the basis for many of our examples in this chapter. Data simulated from this model is shown in Figure 4.1 for two values of β. It can be seen that both z_{2t} series are scaled versions of z_{1t}, perturbed with noise. (In the simulation we generated 300 $\mathbf{a}_t$s with $\Sigma_a=\begin{bmatrix}1.0 & \\ 0.8 & 1.0\end{bmatrix}$ and filtered these through equations [4.1] for both values of β, starting with $z_{1,t-1}=0$. The last 100 values of each series were used in order to minimise startup effects.) This is of course an artificially simple example, however, such redundancy may often occur if the system under study is poorly understood (as can be the case in, for example, economic situations).

4.2.1 VARMA models for system [4.1]

This section investigates the various VARMA models which could be used to describe the system [4.1] and the effects which the presence of CI has upon them. We can write model [4.1] in VARMA notation (with a contemporaneous coefficient matrix) as

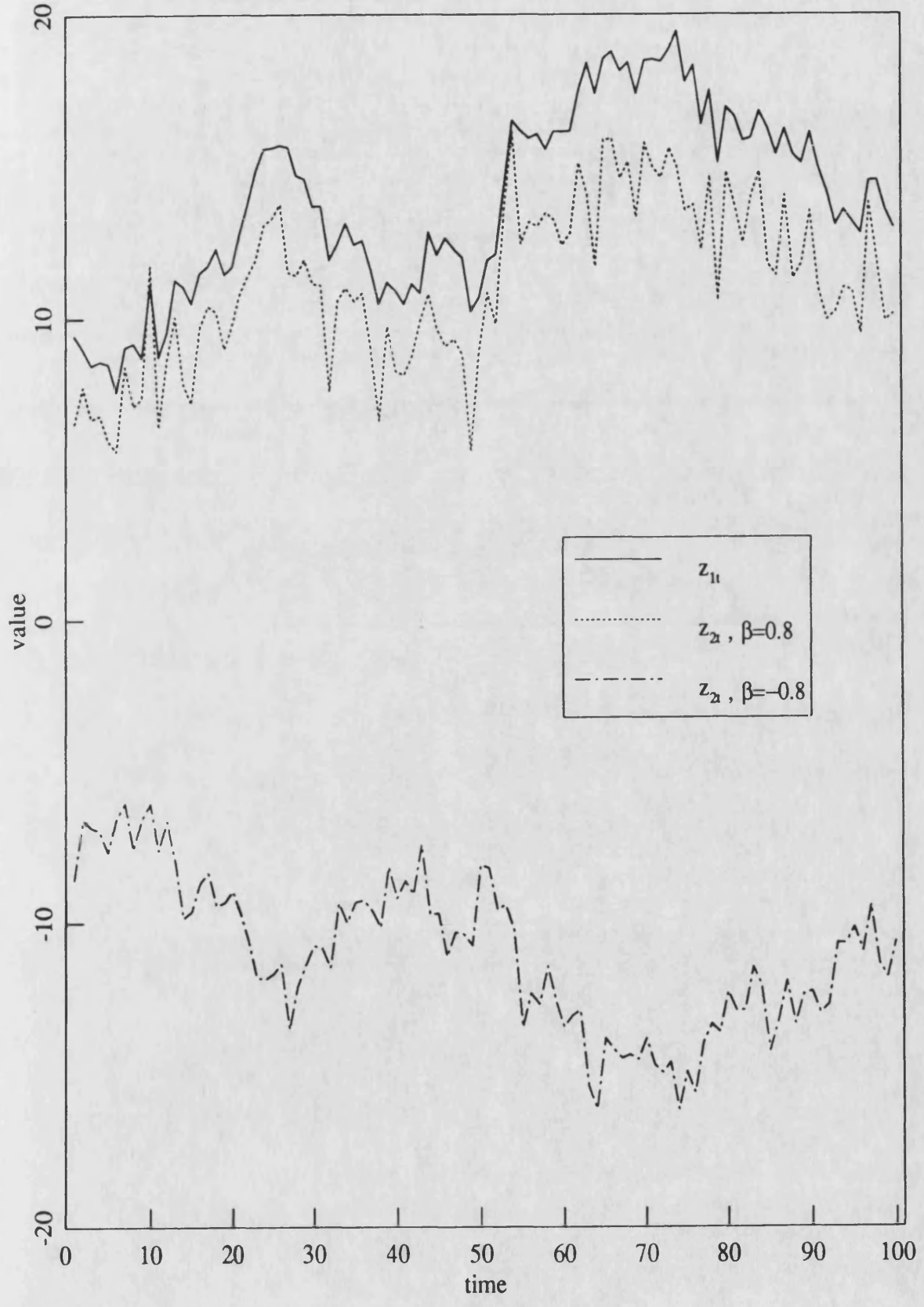
$$\left(T-\Phi_1^{[2]}B\right)\mathbf{z}_t=\mathbf{a}_t, \qquad [4.2]$$

where $T=\begin{bmatrix}1 & 0\\ -\beta & 1\end{bmatrix}$ and $\Phi_1^{[2]}=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$, so that $\Phi^{[2]}(B)=\begin{bmatrix}1-B & 0\\ -\beta & 1\end{bmatrix}$. This is not in the usual form [2.2] for a VARMA model, but it can be put into this form by premultiplying by $T^{-1}=\begin{bmatrix}1 & 0\\ \beta & 1\end{bmatrix}$, to give

$$\left(I-\Phi_1^{[3]}B\right)\mathbf{z}_t=\mathbf{b}_t, \qquad [4.3]$$

where $\Phi_1^{[3]}=T^{-1}\Phi_1^{[2]}=\begin{bmatrix}1 & 0\\ \beta & 0\end{bmatrix}$, so that $\Phi^{[3]}(B)=\begin{bmatrix}1-B & 0\\ -\beta B & 1\end{bmatrix}$, and $\mathbf{b}_t=T^{-1}\mathbf{a}_t$. This model includes a lagged value of z_{1t} and adjusts the second noise term accordingly. The determinantal polynomial of the AR operator of [4.3] ($|\Phi^{[3]}(B)|=1-B$) has one zero on the unit circle, so that Tiao and Box's (1981) suggestion (see §3.1.2) would be sensible to handle this. However, as mentioned in §3.1.2, the estimation routine based upon the state space representation of the VARMA model does not allow such unit roots. Another way to obtain a representation without the contemporaneous coefficient matrix is to factorise the T matrix out of [4.2] to give a model in terms of the transformed data $T\mathbf{z}_t$,

Figure 4.1 Data simulated from [4.1]



$$\left(I-\Phi_1^{[4]}B\right)\mathbf{y}_t=\mathbf{a}_t, \qquad [4.4]$$

where $\Phi_1^{[4]}=\Phi_1^{[2]}T^{-1}=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$, so $\Phi^{[4]}(B)=\begin{bmatrix}1-B & 0\\ 0 & 1\end{bmatrix}$, and $\mathbf{y}_t=T\mathbf{z}_t$. The AR determinantal polynomial of this model is also $1-B$, but only one component of $\mathbf{y}_t$ is non–stationary. We can difference this component to eliminate the unit root, resulting in

$$\mathbf{y}_t^{[5]}=(\nabla y_{1t},y_{2t})^T=(\nabla z_{1t},z_{2t}-\beta z_{1t})^T=\mathbf{a}_t, \qquad [4.5]$$

which is of course the natural representation of the relationship in [4.1] as well as being well suited to estimation (trivial in this case). Notice that it involves one CI term and one differenced term. In most systems we would like to arrive at the equivalent version of this model (for simplicity of estimation) and in §4.3 and §4.4 we present the necessary tools to find and test for such CI relationships.

4.2.2 Overdifferencing

If we were to difference both of the series (as we might be tempted to do, since they are both n.s.), then the differenced series would follow models

$$\nabla z_{1t}=a_{1t} \qquad [4.6a]$$

$$\nabla z_{2t}=\beta a_{1t}+\nabla a_{2t}, \qquad [4.6b]$$

which could be collected together into a VARMA form

$$\nabla\mathbf{z}_t=\left(T^{-1}-\Theta_1^{[7]}B\right)\mathbf{a}_t \text{ , where } \Theta_1^{[7]}=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}. \qquad [4.7]$$

Again, to remove the contemporaneous matrix T^{-1} we can premultiply by T to give

$$\nabla\mathbf{y}_t=\left(I-\Theta_1^{[8]}B\right)\mathbf{a}_t, \qquad [4.8]$$

where $\Theta_1^{[8]}=T\Theta_1^{[7]}=\Theta_1^{[7]}$, $\Theta^{[8]}(B)=\begin{bmatrix}1 & 0\\ 0 & 1-B\end{bmatrix}$, or we can factorise out T^{-1} giving

$$\nabla\mathbf{z}_t=\left(I-\Theta_1^{[9]}B\right)\mathbf{b}_t, \qquad [4.9]$$

with $\Theta_1^{[9]}=\Theta_1^{[7]}T=\begin{bmatrix}0 & 0\\ -\beta & 1\end{bmatrix}$, $\Theta^{[9]}(B)=\begin{bmatrix}1 & 0\\ -\beta B & 1-B\end{bmatrix}$. Now in [4.8] and [4.9] we have induced a unit root in each of the MA parts by differencing both of the component series. This is clearly not satisfactory and will lead to problems in the model specification (E&G), estimation (see §3.1.2) and later analysis.

In the presence of CI, we seek to establish the CI relationships in order to be able

to reduce each series to stationarity (which we require for estimation), but without inducing unit roots in the MA polynomial, which would create more problems. In Table 4.1 we summarise the features of the various models which we have considered for our CI system.

Table 4.1 Summary of the features of VARMA models for [4.1].

Model	(p,q)	contemp?	Data	Staty?	Noise	$\lvert\Phi(B)\rvert$	$\lvert\Theta(B)\rvert$
[4.2]	(1,0)	Φ_0	$\mathbf{z}_t$	NO	$\mathbf{a}_t$	$1-B$	1
[4.3]	(1,0)	NO	$\mathbf{z}_t$	NO	$\mathbf{b}_t$	$1-B$	1
[4.4]	(1,0)	NO	$\mathbf{y}_t$	NO	$\mathbf{a}_t$	$1-B$	1
[4.5]	(0,0)	NO	$\mathbf{y}_t^{[5]}$	YES	$\mathbf{a}_t$	1	1
[4.7]	(0,1)	Θ_0	$\nabla\mathbf{z}_t$	YES	$\mathbf{a}_t$	1	$1-B$
[4.8]	(0,1)	NO	$\nabla\mathbf{y}_t$	YES	$\mathbf{a}_t$	1	$1-B$
[4.9]	(0,1)	NO	$\nabla\mathbf{z}_t$	YES	$\mathbf{b}_t$	1	$1-B$

From this table we see that only representation [4.5] is suitable for estimation, since it has no unit roots in either polynomial and uses stationary data – it is thus the one which we would prefer to work with.

4.2.3 Modelling

If s of our k component series are stationary (but the remainder are n.s.), the stationary series will be excluded from consideration when searching for CI. If we then find r linearly independent CI vectors (collected into a matrix $\underset{(k-s)\times r}{\boldsymbol{\alpha}}$), we can analyse the stationary data $D(B)TR\mathbf{z}_t$, where

$$T=\begin{array}{|c|c|c|}\hline I_s & \underset{s\times(k-s)}{0} & \\ \hline \underset{(k-s)\times s}{0} & \underset{r\times(k-s)}{\boldsymbol{\alpha}^T} & \\ \hline & \underset{(k-s-r)\times r}{0} & I_{k-s-r}\\ \hline\end{array} \qquad D(B)=\begin{array}{|c|c|}\hline I_{s+r} & \underset{(s+r)\times(k-s-r)}{0}\\ \hline \underset{(k-s-r)\times(s+r)}{0} & \nabla I_{k-s-r}\\ \hline\end{array}$$

and $\underset{k\times k}{R}$ reorders the components to put the s stationary ones first. This method of reducing all of the series to stationarity uses the "minimal" amount of differencing and will not induce any unit roots in the MA polynomial. An example of such transformation and differencing is given in §4.5.1.

4.3 Finding possible CI vectors

E&G propose a two–stage testing procedure which involves finding candidate CI vectors, $\boldsymbol{\alpha}_i$ (for which $x_{it}=\boldsymbol{\alpha}_i^T\mathbf{z}_t$ are nearest to stationarity), and then testing the resulting x_{it} to determine if they can be taken to be stationary, in which case $\boldsymbol{\alpha}_i$ defines a CI relationship. In this section we derive and discuss their method for finding $\boldsymbol{\alpha}_i$, exploring many practical details which were not considered by E&G (particularly concerning the CI rank). In §4.4 we consider the tests for stationarity of x_{it} which they presented, demonstrating the application of these tests on some simulated examples and we propose a strategy for finding CI vectors in §4.4.2.

4.3.1 Single CI vector

Let us suppose that there exists only one CI vector, so that in this case we seek a vector $\boldsymbol{\alpha}$ such that $x_t=\boldsymbol{\alpha}^T\mathbf{z}_t$ is as near to stationarity as possible. E&G call x_t the **equilibrium error**, which would be zero in the presence of perfect CI. To find such linear combinations we could minimise this error over possible $\boldsymbol{\alpha}$, i.e. $\min_{\boldsymbol{\alpha}}\sum_t x_t^2$ (the usual least squares form which is an intuitively sensible approach and is shown by E&G to give rise to (a) a relatively simple theory and (b) "well–behaved" estimates. Other possibilities, such as $\sum|x_t|$ will make the analysis more complicated and may not give better results than the least squares solution). The problem is to minimise the variance of x_t,

$$\sum_{t=1}^{N}(\boldsymbol{\alpha}^T\mathbf{z}_t)^2=\sum_{t=1}^{N}\boldsymbol{\alpha}^T\mathbf{z}_t\mathbf{z}_t^T\boldsymbol{\alpha}=\boldsymbol{\alpha}^T M\boldsymbol{\alpha}, \quad (M=\sum_{t=1}^{N}\mathbf{z}_t\mathbf{z}_t^T) \qquad [4.10]$$

(where N is the number of observations available on $\mathbf{z}_t$) subject to an identification constraint. Notice that $\hat{\boldsymbol{\alpha}}=\underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\sum_t x_t^2$ does not necessarily give us a stationary series, x_t, but E&G state that asymptotically all linear combinations of $\mathbf{z}_t$ (with n.s. components) have infinite variance, except the CI one (if it exists), so that CI vectors will be (approximately) given by $\hat{\boldsymbol{\alpha}}$.

In order to use the regression results (to be described below), E&G suggest constraining a coefficient to be equal to 1 (although we could consider other constraints – see e.g. §4.3.2). This constraint takes the form

$$\boldsymbol{\alpha}=\mathbf{q}+Q\boldsymbol{\theta}, \text{ where } \underset{k\times 1}{\mathbf{q}}=\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}, \underset{k\times(k-1)}{Q}=\begin{bmatrix}0\cdots 0\\ I_{(k-1)}\end{bmatrix} \text{ and } \underset{(k-1)\times 1}{\boldsymbol{\theta}}=\begin{bmatrix}\theta_2\\\vdots\\\theta_k\end{bmatrix}, \qquad [4.11]$$

i.e. $\boldsymbol{\alpha}=(1,\theta_2,\ldots,\theta_k)^T$ (assuming that the first variable is included in the CI vector with non–zero coefficient – we can reorder the components if this is not the case).

The solution to minimising [4.10] subject to the constraints [4.11] is given by

$$\min_{\alpha \text{ s.t. } [4.11]} \alpha^T M \alpha \iff \min_{\theta} (\mathbf{q}+Q\theta)^T M (\mathbf{q}+Q\theta).$$

Expanding this second expression, differentiating with respect to θ and equating to zero gives

$$\hat{\theta} = -(Q^T M Q)^{-1} Q^T M \mathbf{q} \text{ and so}$$

$$\hat{\alpha} = \mathbf{q} - Q(Q^T M Q)^{-1} Q^T M \mathbf{q}.$$

If we now consider the regression of z_{1t} on $z_{2t},\ldots,z_{kt}$, in the usual notation (with no intercept term) we have

$$\mathbf{y} = \begin{bmatrix} z_{11} \\ \vdots \\ z_{1N} \end{bmatrix}, \text{ and } X = \begin{bmatrix} z_{21} & \cdots & z_{k1} \\ & \vdots & \\ z_{2N} & \cdots & z_{kN} \end{bmatrix} \qquad [4.12]$$

so that $M = \begin{bmatrix} \mathbf{y}^T\mathbf{y} & \mathbf{y}^T X \\ X^T\mathbf{y} & X^T X \end{bmatrix}$, i.e. $Q^T M Q = X^T X$ and $Q^T M \mathbf{q} = X^T \mathbf{y}$ which gives

$$\hat{\alpha} = \mathbf{q} - Q(X^T X)^{-1} X^T \mathbf{y} = \mathbf{q} - Q\hat{\beta},$$

where $\hat{\beta} = (\hat{\beta}_2,\ldots,\hat{\beta}_k)^T$ = the ordinary least squares estimate of the coefficients in the regression $\mathbf{y} = X\beta + \varepsilon$, i.e. $\hat{\alpha} = (1, -\hat{\beta}_2,\ldots,-\hat{\beta}_k)^T$. Notice also that the estimated equilibrium errors are given by $(\hat{x}_1,\ldots,\hat{x}_N)^T = \hat{\varepsilon}$. Thus the constraints [4.11] have led us to a relatively simple solution which can be shown to have good properties – E&G quote a theorem due to Stock which proves the consistency and convergence properties of the estimates. Although many regression assumptions are violated and such regressions have been called "spurious" by Granger and Newbold (1986) when the series are **not** CI (and there is no vector α to be estimated, in which case the results are misleading and the consistency etc. properties fail to hold), many useful results hold when CI is present. For example, E&G state and prove a theorem establishing that, under CI, the least squares standard errors of the estimates are consistent estimates of the true standard errors, which we can use to obtain better estimates of the CI vector(s) (see example (b) below). CI variables will be highly collinear (by their nature) which may cause problems when developing models (due to the near non–singularity of the matrix $X^T X$). Wetherill (1986, Chapter 4), for example covers many techniques to help when this is the case, including the use of principal components – see §4.3.2.

Examples

We now consider the application of this regression approach to some simulated systems (based on [4.1]) which are CI.

(a) $k=2$

First consider the example system of [4.1], upon which we can perform two possible regressions

$$z_{1t}=\alpha_1 z_{2t}+\varepsilon_{1t}, \qquad [4.13a]$$

$$z_{2t}=\alpha_2 z_{1t}+\varepsilon_{2t}. \qquad [4.13b]$$

With data simulated from model [4.1] (with $\beta=0.8$), regression [4.13b] gives $\hat{\alpha}_2=0.79\simeq\beta$ (standard error=0.006, assuming the series to be CI) while [4.13a] gives $\hat{\alpha}_1=1.25=1/0.8\simeq 1/\beta$ (s.e.=0.01). It can be shown that, in general, $\hat{\alpha}_1=\hat{\rho}^2/\hat{\alpha}_2$ where $\hat{\rho}$ is the estimated instantaneous correlation between series z_{1t} and z_{2t}. In the presence of CI, since the variables are n.s. but "trending" together in some way, the correlation approaches one in large samples (in the simulation $\hat{\rho}=0.93$), so $\alpha=(-\beta,1)^T$ is estimated well by either of the regressions which we can choose. Notice that we only need to perform one of the regressions to obtain this estimate since the CI relationship (if it exists) **must** include both (n.s.) series with non–zero coefficients. Having obtained an estimate of the "best" CI vector (i.e. the vector which gives us a series nearest to stationarity) we then need to test the residuals ε_{1t} (or ε_{2t}) to establish whether the vector actually represents a CI relationship – §4.4 deals with such testing.

(b) $k>2$

E&G only illustrated the case when $k=2$ (or effectively two, since they only considered pairs of variables for their examples with larger k). When $k>2$, to determine the single CI vector we need to perform at most $k-1$ regressions – z_{it} on $\mathbf{z}_{-i,t}$ $(=(z_{1t},\ldots,z_{i-1,t}z_{i+1,t},\ldots,z_{kt})^T$, $\mathbf{z}_t$ excluding the component i) for $i\in$ any $(k-1)$ subset of $\{1,\ldots,k\}$ (since the CI linear combination has to include at least two variables). To demonstrate this we extend Box and Tiao's (1977) example model [4.1] to include

$$z_{3t}=z_{3,t-1}+a_{3t}, \quad \mathbf{a}_t\sim N_3(\mathbf{0},\Sigma_a). \qquad [4.1c]$$

The trivariate system $\mathbf{z}_t$ is still CI, with single CI vector $\alpha=(-\beta,1,0)^T$, but the third component is not included in the CI relationship. There are three possible regressions we could perform, with each of the components in turn as dependent variable

$$z_{1t}=\alpha_2^{[1]}z_{2t}+\alpha_3^{[1]}z_{3t}+\varepsilon_{1t}, \qquad [4.14a]$$

$$z_{2t}=\alpha_1^{[2]}z_{1t}+\alpha_3^{[2]}z_{3t}+\varepsilon_{2t}, \qquad [4.14b]$$

$$z_{3t}=\alpha_1^{[3]}z_{1t}+\alpha_2^{[3]}z_{2t}+\varepsilon_{3t}. \qquad [4.14c]$$

From regression [4.14b] we obtain (from the simulation with $\beta=0.8$, $\Sigma_a=\begin{bmatrix}1.0 & & \\ 0.8 & 1.0 & \\ 0 & 0 & 1.0\end{bmatrix}$) $\hat{\alpha}^{[2]}=(-\alpha_1^{[2]},1,-\alpha_3^{[2]})^T=(-0.79,\ 1,\underset{(0.012)}{0.003})^T\approx\alpha$ (s.e. in brackets below, assuming that the series ε_{2t} is stationary). A better estimate of α can now be obtained by regressing z_{2t} on z_{1t} only, since the coefficient $\alpha_3^{[2]}$ is insignificant. This reduced regression gives $\hat{\alpha}^{[2']}=(-0.79,\ 1,\ 0)^T$ (c.f. [4.13b] above). [4.14c] will not necessarily find $\hat{\alpha}^{[3]}=\alpha$ since we are constraining $\hat{\alpha}^{[3]}=(-\alpha_1^{[3]},-\alpha_2^{[3]},1)^T$ and z_{3t} is not included in the CI relationship – hence we need to perform at most $k-1$ regressions.

Similarly to the case when $k=2$, the "reverse" (when compared with [4.14b]) regression [4.14a] gives $\hat{\alpha}^{[1]}=(1,-1.21,\underset{(0.015)}{0.026})^T$, which reduces to $\hat{\alpha}^{[1']}=(1,-1.25,\ 0)^T\approx\alpha/(-\beta)$. It can be shown that for the coefficients from the two regressions, $\hat{\alpha}_1^{[2]}\hat{\alpha}_2^{[1]}=\hat{\rho}_{12.3}^2$, where $\hat{\rho}_{12.3}$ is the partial correlation between z_{1t} and z_{2t} neglecting that due to mutual correlation with z_{3t} (see §14.8 of Wei (1990)). With CI variables we would expect this to approach one between variables included in the CI linear combination (in the simulation, $\hat{\rho}_{12.3}=0.98$), so that both [4.14a] and [4.14b] will find $\hat{\alpha}\approx$ some multiple of α. We thus need to determine how many linearly independent CI vectors we have found from the $k-1$ regressions – see §4.3.3.

(c) More complex CI vectors

A different system which exhibits CI can be set up (based on the system [4.1]) using equations [4.1a], [4.1c] and a new version of [4.1b]

$$z_{2t}=\beta z_{1t}+\gamma z_{3t}+a_{2t}. \qquad [4.1b']$$

The CI in this system does not involve only pairs of variables and so cannot be found by regressions on single variables, such as [4.13]. Also the CI vector includes each of the series with non–zero coefficient, so each of the regressions [4.14] should find it. Simulating from this new system of equations (with $\beta=0.8$ and $\gamma=-0.6$, i.e. $\alpha=(-0.8,1,0.6)^T$) and applying the three regressions in [4.14], we find

[4.14a]: $\hat{\alpha}^{[1]T}=(1,-1.21,-0.75)=-1.21(-0.83,1,0.62)$,

[4.14b]: $\hat{\alpha}^{[2]T}=(-0.79,1,0.60)$,

[4.14c]: $\hat{\alpha}^{[3]T}=(-1.32,1.61,1)=1.61(-0.82,1,0.62)$.

Also $\hat{\rho}_{12.3}=0.98$, $\hat{\rho}_{13.2}=0.93$ and $\hat{\rho}_{23.1}=-0.98$, so that the "reverse" regressions each nearly give us the desired vector as we see from the simulation. Again we

need to decide how many CI relationships these vectors describe – in this case they can be seen to be approximately multiples of each other.

(d) Stationary series

Now we consider the case when a stationary series is used in the CI analysis. For this we use [4.1a], [4.1b] and the stationary series $z_{3t}=a_{3t}$. The three possible regressions on the simulated data give us (with standard errors in brackets below)

[4.14a]: $\hat{\alpha}^{[1]T}=(1, \underset{(0.01)}{-1.26}, \underset{(0.13)}{0.13}) \Rightarrow (1, -1.25, 0)$, (excluding the third component)

[4.14b]: $\hat{\alpha}^{[2]T}=(\underset{(0.007)}{-0.79}, 1, \underset{(0.10)}{-0.12}) \Rightarrow (0.79, 1, 0)$,

[4.14c]: $\hat{\alpha}^{[3]T}=(\underset{(0.08)}{0.08}, \underset{(0.09)}{-0.12}, 1) \Rightarrow (0,0,1)$,

which are the same as we would get by not including z_{3t} (regression [4.14c] simply finds the stationary series), suggesting that inclusion of a stationary series does not affect the results. However, it is not clear whether this will hold in general, so to avoid any possible problems it is best to exclude any stationary series from consideration when searching for CI, as suggested earlier.

4.3.2 Principal components

An alternative constraint to [4.11] which might appear to be more natural is $\alpha^T\alpha=1$ which (using a result from multivariate analysis) gives $\hat{\alpha}$=argmin [4.10] subject to $\alpha^T\alpha=1$ to be the eigenvector corresponding to the smallest eigenvalue of M, i.e. $\hat{\alpha}$ is the k'th **principal component** of the data ([**y**:X] of [4.12]; see e.g. Chatfield and Collins (1980)). E&G suggest that there may be other consistent estimates of the CI vector and since principal components are chosen to be linearly independent, this may be a fruitful route to pursue which avoids any difficulties with "reverse" regressions and directly estimates the CI rank, r for multiple CI vectors. We do not have any analogous results to those of E&G to suggest that the principal component estimates are consistent, however, it may be a useful preliminary analysis to perform in order to estimate r and perhaps even the CI vectors where it might highlight the collinearities among the variables and suggest transformations/combinations to use in order to avoid problems.

Example

Continuing with our system [4.1a], [4.1b'] and [4.1c] from example (c) above, we can perform a principal components analysis upon the system in search of CI vectors (linear combinations with small variance – i.e. nearest to stationary), which gives the following.

component	proportion of variance	vector
1	0.861	(0.06, −0.48, 0.88)
2	0.134	(0.80, 0.55, 0.24)
3	0.004	(−0.60, 0.68, 0.41)

This nearly finds the CI vector ($\hat{\alpha}=0.68(-0.88,1,0.6)^T$) from the third component, although the estimate is not as good as that obtained from the regressions. We can choose a proportion of the variance to be accounted for (perhaps 0.95 or 0.99) beyond which the vectors are taken to be candidate CI vectors – we can then test the transformed components using the techniques of §4.4.

4.3.3 CI rank

One useful application of principal components to CI is to determine the CI rank r of the system. For example, for the system given by [4.1a], [4.1b'] and [4.1c] we obtain the vectors given in §4.3.1(c) above as possible CI vectors after running regressions [4.14]. Assuming that the tests to be outlined in §4.4 suggest that each of these is a CI vector we then need to find how many linearly independent vectors they actually define. To do this we can perform a principal components analysis on them. For the simulation we only find one significant component (which accounts for 0.99996 of the variance), confirming that in this case $r=1$. One of the results from the regressions will provide the best estimate of the CI vector.

4.3.4 Multiple CI vectors

In the case of multiple CI vectors: $\alpha=[\alpha_1 : \cdots : \alpha_r]$, we can choose α to minimise the sum of squared residuals from equilibrium, which now takes the form

$$\sum_{j=1}^{r}\sum_{t=1}^{N}\alpha_j^T \mathbf{z}_t \mathbf{z}_t^T \alpha_j = \sum_{j=1}^{r}\alpha_j^T M \alpha_j = \mathrm{tr}(\alpha^T M \alpha). \qquad [4.15]$$

In order to obtain similar regression results as the solution to this minimisation we can generalise the constraint [4.11] above in the following way (with the components of $\mathbf{z}_t$ reordered to ensure that the first r are included in the CI linear combinations)

$$\mathrm{vec}\alpha = \mathbf{q} + Q\theta, \text{ where} \qquad [4.16]$$

$$\underset{kr\times 1}{\mathbf{q}} = (1,0,\ldots,0,\ 0,1,0,\ldots,0,\ \ldots\ldots,\ 0,\ldots,0,\underset{(r)}{1},\ldots,0)^T,$$

$$\underset{(k-1)r\times 1}{\theta} = [\theta_{12},\ldots,\theta_{1k},\ \theta_{21},\theta_{23},\ldots,\theta_{2k},\ \ldots\ldots,\ \theta_{r1},\ldots,\theta_{r,r-1},\theta_{r,r+1},\ldots,\theta_{r,k}]^T \text{ and}$$

$$\underset{kr\times(k-1)r}{Q} = \begin{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \vdots \\ & & \ddots & \\ \vdots & & & 1\ 0 \\ 0 & \cdots & & 0\ 1 \end{bmatrix} & & & 0 \\ & \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \vdots \\ & & \ddots & \\ \vdots & & & 1\ 0 \\ 0 & \cdots & & 0\ 1 \end{bmatrix} & & \\ & & \ddots & \\ 0 & & & \begin{matrix} (r)\rightarrow \end{matrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \vdots \\ & & \ddots & \\ 0 & 0 & \cdots & \cdots\ 0 \\ & & & \ddots \\ & & & 1\ 0 \\ 0 & \cdots & & 0\ 1 \end{bmatrix} \end{bmatrix}$$

Each diagonal block in Q is a $k\times(k-1)$ matrix, constructed by inserting a row of zeroes at various places in I_{k-1} (the remainder of Q consists of 0s) so that this simply picks out the components of $\boldsymbol{\theta}$ and puts them into the relevant zero positions in $\mathbf{q}$ – constraining one component of $\boldsymbol{\alpha}_i$ to 1 in each of the r CI vectors.

E&G state the results given in [4.17] and [4.18] below and the regression solution, but we will derive them fully here. To minimise [4.15] subject to the constraints [4.16], we can rewrite

$$\begin{aligned} \mathrm{tr}(\alpha^T M\alpha) &= \mathrm{tr}(\alpha^T A^T A\alpha) && (M=A^TA) \\ &= (\mathrm{vec}(A\alpha))^T \mathrm{vec}(A\alpha) && (\mathrm{tr}(B^TB)=(\mathrm{vec}B)^T\mathrm{vec}B) \\ &= ((I_r\otimes A)\mathrm{vec}\alpha)^T(I_r\otimes A)\mathrm{vec}\alpha && (\mathrm{vec}(ABC)=(C^T\otimes A)\mathrm{vec}B) \\ &= (\mathrm{vec}\alpha)^T(I_r\otimes A^T)(I_r\otimes A)\mathrm{vec}\alpha && ((A\otimes B)^T=A^T\otimes B^T) \\ &= (\mathrm{vec}\alpha)^T(I_r\otimes A^TA)\mathrm{vec}\alpha && ((A\otimes B)(C\otimes D)=(AC\otimes BD)) \\ &= (\mathrm{vec}\alpha)^T(I_r\otimes M)\mathrm{vec}\alpha \end{aligned}$$

so that

$$\min_{\alpha \text{ s.t. } [4.16]} \mathrm{tr}(\alpha^T M\alpha) \iff \min_{\alpha \text{ s.t. } [4.16]} (\mathrm{vec}\alpha)^T(I_r\otimes M)\mathrm{vec}\alpha$$

$$\iff \min_{\boldsymbol{\theta}} (\mathbf{q}+Q\boldsymbol{\theta})^T(I_r\otimes M)(\mathbf{q}+Q\boldsymbol{\theta}). \qquad [4.17]$$

Expanding this expression, differentiating with respect to $\boldsymbol{\theta}$ and equating to zero gives the solution for $\boldsymbol{\theta}$ given in E&G, i.e.

$$\mathrm{vec}\hat{\alpha}=\mathbf{q}-Q(Q^T(I_r\otimes M)Q)^{-1}Q^T(I_r\otimes M)\mathbf{q}. \qquad [4.18]$$

In a multiple regression of z_{it} on $\mathbf{z}_{-i,t}$ (see §4.3.1(b)) we have

$$\mathbf{y}_i=\begin{bmatrix} z_{i1} \\ \vdots \\ z_{iN} \end{bmatrix} \text{ and } X_{-i}=\begin{bmatrix} \mathbf{z}_{-i,1}^T \\ \vdots \\ \mathbf{z}_{-i,N}^T \end{bmatrix}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_i=(X_{-i}^TX_{-i})^{-1}X_{-i}^T\mathbf{y}_i=(M_{-i,-i})^{-1}M_{-i,i}$$

($M_{-i,-i}$ denotes the matrix M with row and column i removed, $M_{-i,i}$ column i of M with row i removed). Now $Q^T(I_r\otimes M)Q$ can be shown to be

$$\begin{bmatrix} M_{-1,-1} & & & \\ & M_{-2,-2} & & \\ & & \ddots & \\ & & & M_{-r,-r} \end{bmatrix}$$

since postmultiplication of a block diagonal matrix by Q removes columns from the blocks and thus premultiplication by Q^T removes rows from the blocks. Similarly it can be shown that $Q^T(I_r\otimes M)\mathbf{q}=\begin{bmatrix} M_{-1,1} \\ \vdots \\ M_{-r,r} \end{bmatrix}$ (since postmultiplication of a block diagonal matrix by $\mathbf{q}$ picks out columns from the blocks), so that $\text{vec}\hat{\alpha}=\mathbf{q}-Q\hat{\beta}$ where $\hat{\beta}=(\hat{\beta}_1^T,\hat{\beta}_2^T,\ldots,\hat{\beta}_r^T)^T$ and the joint solution of [4.15] subject to [4.16] is equivalent to performing the regressions z_{it} on $\mathbf{z}_{-i,t}$, $i=1,..,r$. This derivation assumes that we know the CI rank r, however we can simply perform the $k-1$ regressions (as discussed in §4.3.1 above) and determine the CI rank from these. As in the case of a single CI vector, the consistency properties of the estimates and their standard errors hold when the residual series are stationary (and we have CI relationships).

Example

To illustrate the method for multiple CI vectors, we consider the system defined by [4.1a], [4.1b] and

$$z_{3t}=\gamma z_{1t}+a_{3t}, \qquad [4.1c']$$

which has two CI vectors: $\alpha_1=(-\beta,1,0)^T$ and $\alpha_2=(-\gamma,0,1)^T$. We again simulate from this system, taking $\beta=0.8$, $\gamma=-0.6$, Σ as before. We can perform the three regressions of [4.14], which give us (with standard errors in brackets below)

[4.14a]: $\hat{\alpha}^{[1]T}=(1,\underset{(0.06)}{-0.78},\underset{(0.07)}{0.64})$ which can be shown to be

$$\approx -0.78\alpha_1+0.64\alpha_2,$$

[4.14b]: $\hat{\alpha}^{[2]T}=(\underset{(0.06)}{-0.87},1,\underset{(0.10)}{0.12})$ and deleting the third coefficient

$$\Rightarrow\hat{\alpha}^{[2']T}=(\underset{(0.01)}{-0.79},1,0)\approx\alpha_1,$$

[4.14c]: $\hat{\alpha}^{[3]T}=(\underset{(0.08)}{0.68},\underset{(0.10)}{0.12},1)$ and deleting the second coefficient

$$\Rightarrow\hat{\alpha}^{[3']T}=(\underset{(0.01)}{0.58},0,1)\approx\alpha_2.$$

It is not easy to see that $\hat{\alpha}^{[1]}$ is a linear combination of α_1 and α_2, but having obtained $\hat{\alpha}^{[1]}$, $\hat{\alpha}^{[2']}$ and $\hat{\alpha}^{[3']}$ (and confirmed that they each correspond to a CI relationship) we can perform a principal components analysis to determine r. For these estimated vectors such an analysis suggests that $r=2$ since the first two principal components account for all of the variance between the vectors (first alone 0.93). Any two linearly independent vectors from the regressions can thus be chosen to represent CI relationships. In this case it is easy to see that $\hat{\alpha}^{[2']}$ and $\hat{\alpha}^{[3']}$ are linearly independent.

4.4 Testing

Having found suitable candidate vector(s) α_i, from the regression analyses described above, we need to test the residuals for stationarity to see if the vector actually defines a CI relationship. The tests given in E&G are all versions of some of the usual tests for stationarity, however in the context of CI we are testing the null hypothesis that the series are **not** CI and hence that the residuals are still n.s. We will only discuss the simpler tests recommended by E&G, although sometimes the more complex tests may be required. A simple and useful guide in some cases will be to look at the acf of x_{it} with bounds of $2/\sqrt{N}$ – see some of the examples which follow.

4.4.1 Tests

(DW) Durbin–Watson

This test is based on the following statistic

$$\hat{DW}=\frac{\sum_{t=2}^{N}(\hat{x}_{it}-\hat{x}_{i,t-1})^2}{\sum_{t=1}^{N}\hat{x}_{it}^2}, \qquad [4.19]$$

which can be shown to be $\approx 2(1-\hat{\rho}_1)$ (where $\hat{\rho}_1$ is the acf of $\hat{x}_{it}$ at lag 1). Since $\hat{\rho}_1$ can be positive or negative, we will use the statistics $\hat{DW}$ if $\hat{DW}<2$ ($\hat{\rho}_1>0$) and $4-\hat{DW}$ if $\hat{DW}>2$ ($\hat{\rho}_1<0$). We then have that

$$\hat{DW}\approx 0 \text{ (or } 4-\hat{DW}\approx 0) \Rightarrow \hat{\rho}_1\approx\pm 1 \Rightarrow \hat{x}_{it} \text{ n.s.} \Rightarrow \alpha_i \text{ is not a CI vector,}$$

$$\text{whilst } \hat{DW}\approx 2 \Rightarrow \hat{\rho}_1\approx 0 \Rightarrow \hat{x}_{it} \text{ stationary} \Rightarrow \alpha_i \text{ is a CI vector.}$$

The null hypothesis is of no CI which is not the usual DW test of no autocorrelation in the residuals, so E&G tabulate new critical values for the statistic (from a simulation study, but only using $k=2$, $N=100$; see Table 4.2 below). These values are sensitive to the particular parameters within the null hypothesis (which is composite. E&G distinguish between two cases: (i) The

system is "first order", which defines the null to be $\nabla \mathbf{z}_t = \mathbf{a}_t \sim N_k(\mathbf{0}, \Sigma_a)$ and includes all positive definite covariance matrices Σ_a. However, they show that Σ_a does not affect the tests and therefore if the system can be shown to be first order, the critical values will be those given and the DW test is the most powerful; (ii) The null is simply that the system is stationary in the differences, which includes a full set of VARMA parameters as well. For this case we require an "augmented" test to more fully specify the dynamics. "order" now refers to $\max(p,q)$ of a suitable VARMA model – case (i) should really be called "zeroth order"). The DW test uses only the first acf coefficient and so only tests for first order correlation which may not be satisfactory in certain circumstances.

(DF) Dickey–Fuller

This test uses an auxiliary regression (Dickey and Fuller (1979, 1981)) of the residuals

$$\nabla \hat{x}_{it} = -\varphi \hat{x}_{i,t-1} + \varepsilon_t \qquad [4.20]$$

$$\text{or } \hat{x}_{it} = (1-\varphi)\hat{x}_{i,t-1} + \varepsilon_t$$

so that if x_{it} is n.s. (and the system is first order) then $\hat{\varphi} \approx 0$. The test is then on the (modulus of the) t–statistic for the parameter φ, i.e. $\hat{DF} = t_{\varphi}$, although the t distribution is not appropriate. Again, some new critical values are tabulated by E&G (under the null hypothesis of no CI) and these are found to be more stable to differences in the null hypotheses than the values for DW (see Table 4.2).

(ADF) Augmented Dickey–Fuller

This test extends the right hand side of the regression in [4.20] above to allow more dynamics in the model for the residuals (the null hypothesis of [4.20] is that the residuals follow a random walk)

$$\nabla \hat{x}_{it} = -\varphi \hat{x}_{i,t-1} + b_1 \nabla \hat{x}_{i,t-1} + \ldots + b_p \nabla \hat{x}_{i,t-p} + \varepsilon_t, \text{ for some } p, \qquad [4.21]$$

and again the test is the t–statistic of the parameter φ. The null now allows the residuals to follow a p'th order autoregression when differenced, which can provide a better approximation to the true model. We denote the test statistic t_{φ} by $\hat{ADF}(p)$, so that $\hat{ADF}(0) = \hat{DF}$. If the residuals actually follow a random walk, then this test will be inefficient (overparameterised) compared with DF above. However one can always look at the b_i coefficients to decide on how many to include – E&G state that their critical values are insensitive to misspecification of p, but that it is best to only include the significant b_i's.

E&G recommend using ADF (for suitable p), but we will use all three tests (with the quoted critical values as rough guides, although we will rarely be in exactly

the situation for which they were calculated – Engle and Yoo (1987) present some further simulation results for the tests – see §4.7.2) and also examine the acf of the residual series to test for/explore co–integration.

Table 4.2 Critical values for CI tests from Engle and Granger (1987).
(k=2, N=100, 10000 simulations)

	First order case			Fourth order case		
Test	1%	5%	10%	1%	5%	10%
DW	0.511	0.386	0.322	0.455	0.282	0.209
DF	4.07	3.37	3.03	3.90	3.05	2.71
ADF(4)	3.77	3.17	2.84	3.73	3.17	2.91

Examples

We now illustrate the application and results of the tests with our simulated CI system based upon the equations [4.1].

(a)

First we use [4.1a], [4.1b] and [4.1c] which has a single CI vector $\alpha=(-0.8, 1, 0)^T$. The regression analysis for this system was given in §4.3.1(b) above and the results are summarised in the table below. Figures 4.2a–c show the acfs of the residuals ε_{1t}, ε_{2t} and ε_{3t} (together with approximate 95% limits, $\pm 2/\sqrt{N}$), and serve to demonstrate the test results – clearly ε_{1t} and ε_{2t} are stationary, while ε_{3t} is not.

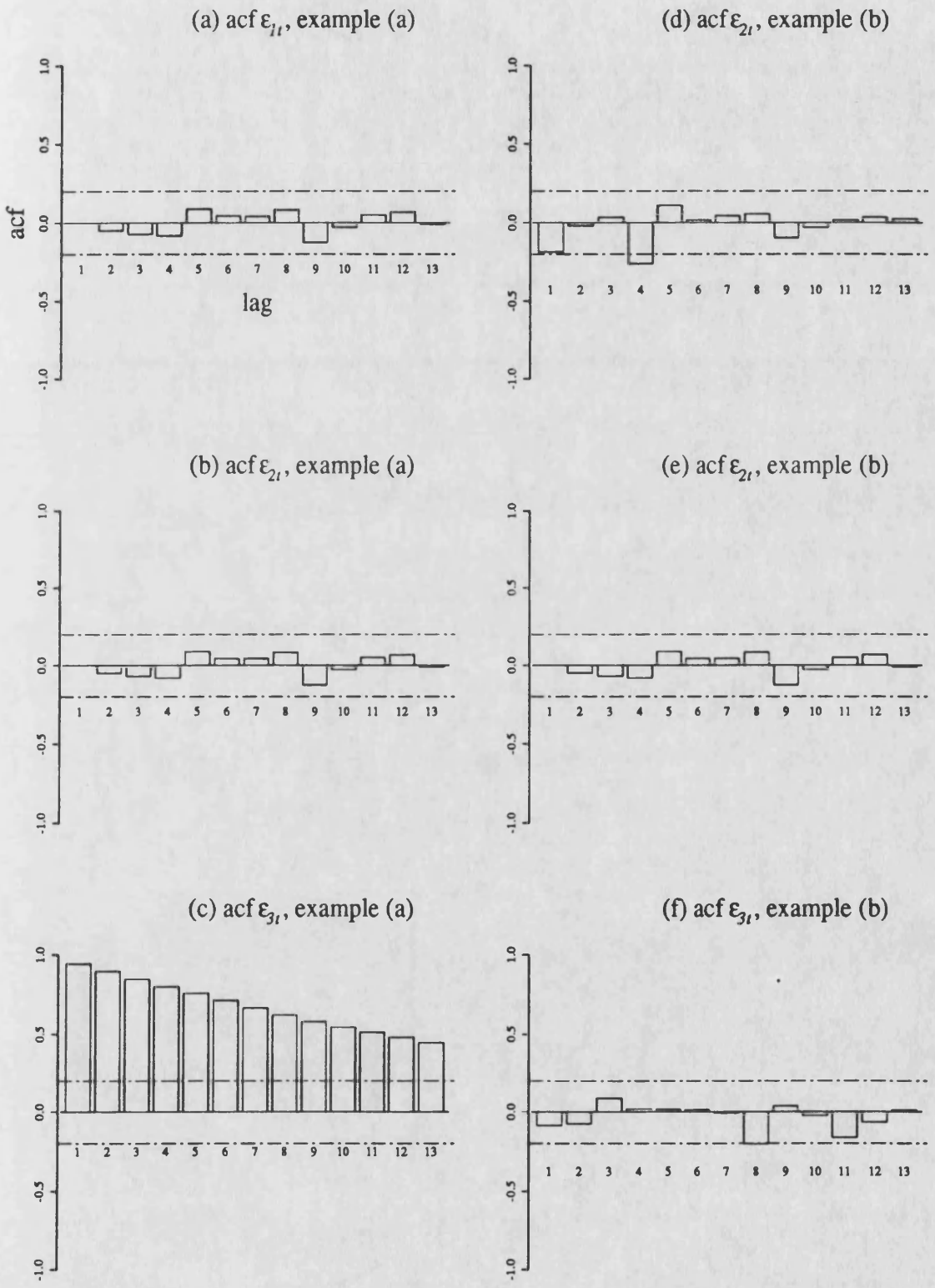
Regression		$\hat{\alpha}$	$D\hat{W}$	$D\hat{F}$	$A\hat{D}F(p)$	Possible CI vector?
[4.14a]	ε_{1t}	(1, −1.25, 0)	1.97	9.77	$\hat{p}$=0	YES
[4.14b]	ε_{2t}	(−0.79, 1, 0)	1.99	9.88	$\hat{p}$=0	YES
[4.14c]	ε_{3t}	(−1.15, 0.17, 1)	0.04	0.68	$\hat{p}$=0	NO

Notice that the test ADF reduces to DF in these cases. The tests suggest that we have found two CI vectors, but the application of principal components analysis to the vectors, as suggested above, reveals that there is really only one (accounting for all of the variance between the vectors).

(b)

A second example uses [4.1a], [4.1b] and [4.1c'] which has two (linearly independent) CI vectors α_1 and α_2. A complete regression analysis was given in §4.3.4 above and the test results are summarised below. Figures 4.2d–f show the acfs of the residuals from the regressions (together with approximate 95% limits, $\pm 2/\sqrt{N}$), which suggest that all three residual series are stationary. These results confirm that all three regressions find CI vectors, but the principal components analysis applied in §4.3.4 found that these could be derived from two linearly

Figure 4.2


(a) acf ε_{1t}, example (a)
(d) acf ε_{2t}, example (b)
(b) acf ε_{2t}, example (a)
(e) acf ε_{2t}, example (b)
(c) acf ε_{3t}, example (a)
(f) acf ε_{3t}, example (b)
acf
lag
1.0
0.5
0.0
-0.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13

Regression	vector	$D\hat{W}$	$D\hat{F}$	$A\hat{D}F(p)$	Possible CI vector?
[4.14a]	$\hat{\alpha}^{[1]}$	1.65	11.87	$\hat{p}=0$	YES
[4.14b]	$\hat{\alpha}^{[2']}$	1.99	9.87	$\hat{p}=0$	YES
[4.14c]	$\hat{\alpha}^{[3']}$	1.84	10.79	$\hat{p}=0$	YES

independent vectors.

4.4.2 Testing strategy

We now summarise the strategy which we can employ to find CI relationships.

1. Regress z_{it} on $\mathbf{z}_{-i,t}$, giving vectors $\hat{\alpha}^{[i]}$ and residuals $\hat{x}_{it}=\hat{\alpha}^{[i]T}\mathbf{z}_t$, for $i\in I\subset\{1,\ldots,k\}$ (I is any $(k-1)$ subset of $\{1,\ldots,k\}$).

2. Test each of the residual series for stationarity using the tests described in §4.4.1 (and other guides such as the acf). If any are stationary (x_{jt}, $j\in J\subseteq I$, say), then the estimates can be improved by repeating the individual regressions z_{jt} on $\mathbf{z}_{-j,t}$, excluding from $\mathbf{z}_{-j,t}$, the variables whose coefficients are judged to be insignificant by their standard errors.

3. The number of CI relationships (CI rank, r) can be determined by performing a principal components analysis on the vectors $\hat{\alpha}^{[j']}$ – the regression results provide the best estimates of the actual CI vectors.

4.5 Case study

To fully illustrate the strategy for exploring and testing for CI in a multivariate time series we will analyse the US Hog data. This dataset is described in §A.4 and has been widely analysed. In Figure 4.3 we sketch the series and their acfs to demonstrate the n.s. of each component.

4.5.1 Analysis

We now perform the k regressions, z_{it} on $\mathbf{z}_{-i,t}$, deleting insignificant coefficients to arrive at suitable candidate CI vectors. The results of this and the tests for stationarity of the residuals, $x_{it}=\hat{\alpha}^{[i]T}\mathbf{z}_{it}$ are summarised in Table 4.3 below and the acfs of the residuals x_{it} are shown in Figures 4.4a–e.

Table 4.3 CI analysis of the Hog data

i	vector					$D\hat{W}$	$D\hat{F}$	$A\hat{D}F(p)$	CI?
1	$\hat{\alpha}^{[1]T}$	1.00	0.34	–0.42	–0.20	–0.32	1.47	7.11	$\hat{p}=0$	YES
2	$\hat{\alpha}^{[2]T}$	1.24	1.00	–0.67	–0.50	–0.74	1.34	7.11	$\hat{p}=0$	YES
3	$\hat{\alpha}^{[3]T}$	–1.11	–0.48	1.00	0.62	0.00	1.31	6.25	$\hat{p}=0$	YES
4	$\hat{\alpha}^{[4]T}$	–1.07	–0.74	1.05	1.00	0.00	1.02	5.79	$\hat{p}=0$	YES
5	$\hat{\alpha}^{[5]T}$	–0.92	–0.71	0.00	0.00	1.00	0.66	4.48	$\hat{p}=0$	YES

(We can choose any 4 regressions from these as discussed in §4.3.1, but we

Figure 4.3 US Hog data

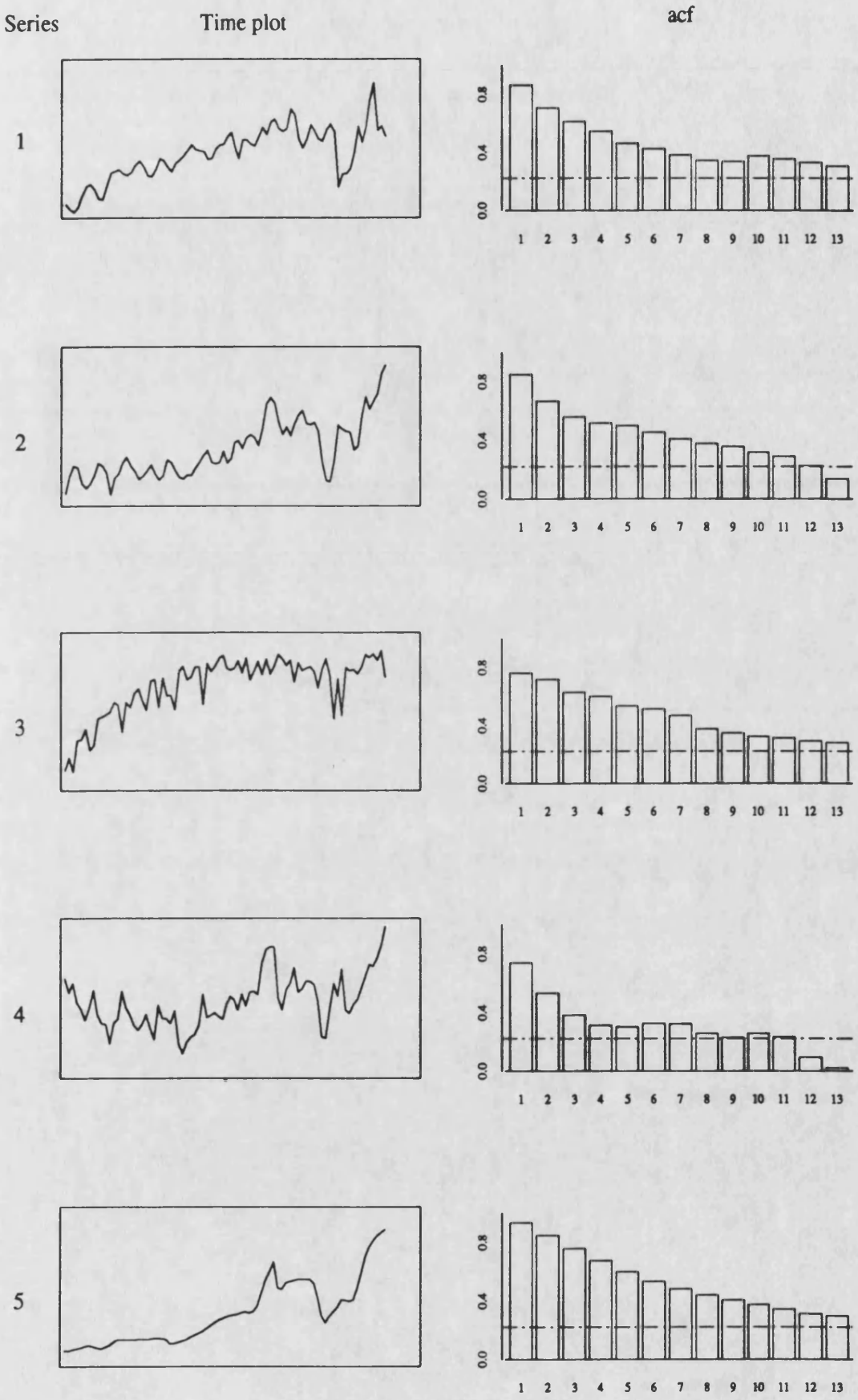
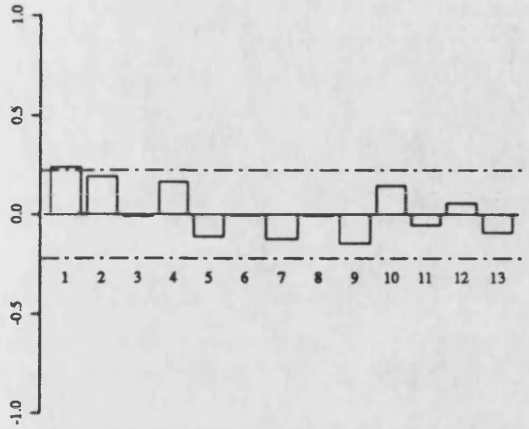
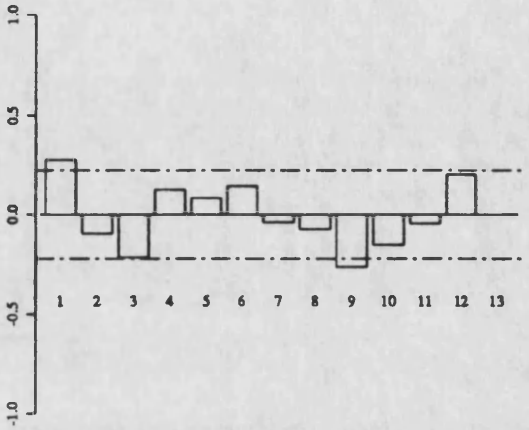


Figure 4.4 acfs of residuals from CI analysis of Hog data

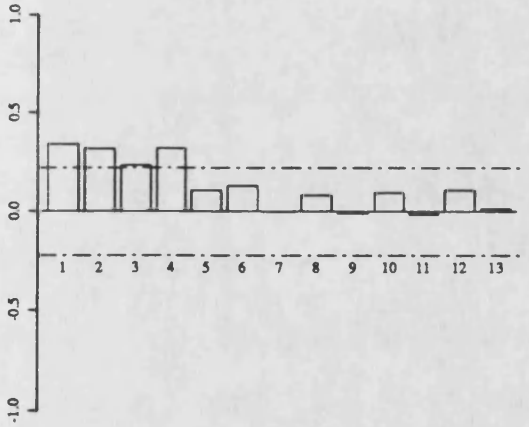
(a) x_{1t}



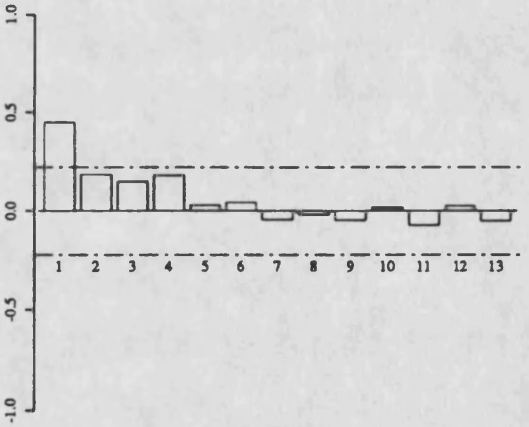
(b) x_{2t}



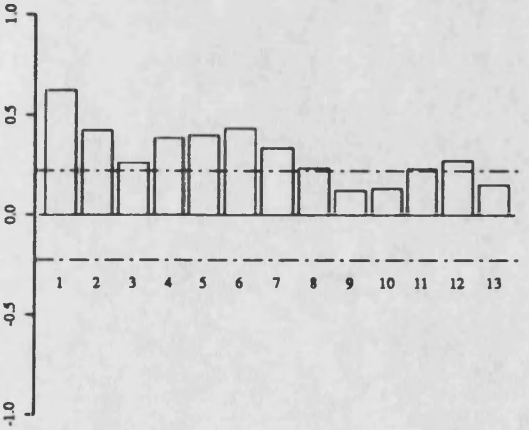
(c) x_{3t}



(d) x_{4t}



(e) x_{5t}



provide all 5 for illustration.) These results suggest that we have found 4 CI vectors (any 4 $\hat{\alpha}^{[i]}$), but a principal components analysis of these reveals that 3 vectors can be found which account for 0.999 of the variation between the 4 vectors. Hence we may choose any 3 linearly independent vectors to represent the CI relationships – for ease of interpretation we have chosen $\hat{\alpha}^{[1]}$, $\hat{\alpha}^{[4]}$ and $\hat{\alpha}^{[5]}$ – so that we can then analyse the stationary series $D(B)T\mathbf{z}_t$, where

$$T=\begin{bmatrix} \hat{\alpha}^{[1]T} \\ \hat{\alpha}^{[4]T} \\ \hat{\alpha}^{[5]T} \\ \begin{bmatrix} 0 & : & I_2 \\ 2\times 3 & & \end{bmatrix} \end{bmatrix}, \quad D(B)=\begin{bmatrix} I_3 & \\ & \nabla I_2 \end{bmatrix},$$

as suggested in §4.2.3. Thus the first three series are given directly by the CI relationships and the remaining two are first differences of, in this case, z_{4t} and z_{5t}. These series are plotted in Figure 4.5 together with their acfs.

In §6.4 we compare the models which we can build for this stationary data and for the original data to see if any benefit can be gained from considering and correcting for the CI in the data.

4.5.2 Interpretation

Although not a primary objective, we can use the definitions of the variables in the Hog data to interpret the actual relationships which we have found. §A.4 defines the data fully, but briefly, the series are each logged and

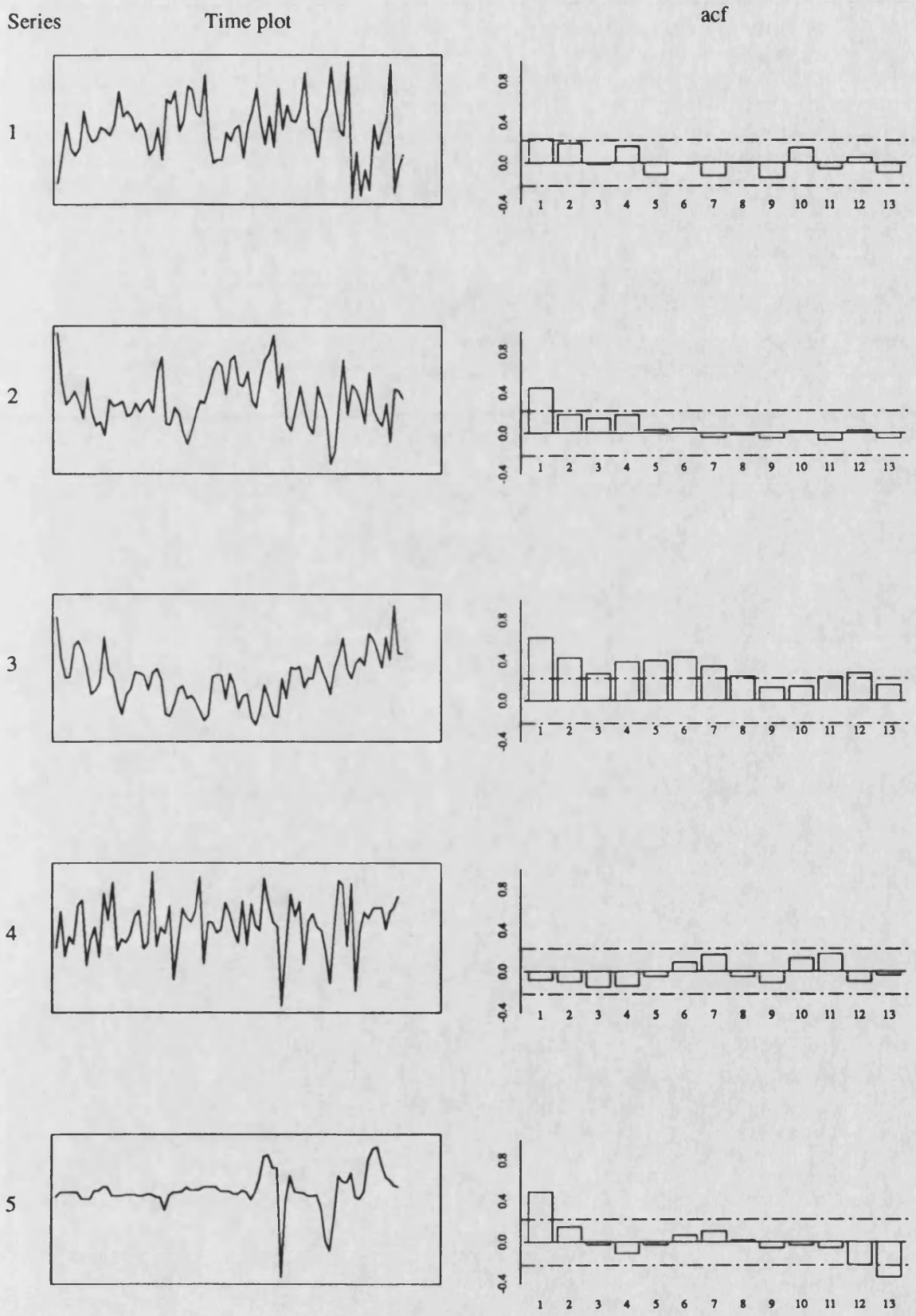
variable	corresponds to	notation	
1	hog numbers	H_n	$=\log h_n$
2	hog price	$H_\$$	$=\log h_\$$
3	corn supply	C_n	$=\log c_n$
4	corn price	$C_\$$	$=\log c_\$$
5	farm wages rate	$W_\$$	$=\log w_\$$

So $\hat{\alpha}^{[1]T}\mathbf{z}_t=H_n+0.34H_\$-0.42C_n-0.20C_\$-0.32W_\$$ approximately corresponds to a ratio (taking antilogs)

$$\left[\frac{h_n^3 h_\$}{(c_n^2 c_\$)^{0.6} w_\$}\right]^{\frac{1}{3}},$$

which represents a balance between a farmers income (from hogs) and outgoings (on their feed and handling) and is thought to be stable from the analysis (the acf of the residuals suggests that it is nearly white noise). $\hat{\alpha}^{[4]}$ suggests that a ratio $\left[\frac{c_n c_\$}{h_n h_\$}\right]$ is stable while $\hat{\alpha}^{[5]}$ corresponds to a stable ratio between the farm wage rate and the farmer's income $\left[\frac{w_\$}{h_n h_\$}\right]$.

Figure 4.5 Transformed, stationary components of Hog data



4.5.3 Other relationships

While the multiple regressions shown above will find linear combinations which are nearest to stationarity and hence nearest to CI, we might be interested in other relationships – for example between pairs of variables, perhaps to avoid problems due to collinearities or to find simpler relationships. The most highly (instantaneously) correlated pair is $(H_{\$}, W_{\$})$ $(\hat{\rho}_{25}=0.94)$. A regression reveals that $H_{\$}-0.97W_{\$}$ appears to be stationary $(\hat{DW}=1.04, \hat{DF}=5.34, \hat{ADF}(1)=7.56\ (\hat{p}=1))$ suggesting a stable ratio between the price of hogs and the farm wage rate $\left[\frac{h_{\$}}{w_{\$}}\right]$, which may be sensible. Another highly correlated pair of variables is (H_n, C_n) $(\hat{\rho}_{13}=0.79)$, and we find that the combination $H_n-0.45C_n$ or $\left[\frac{h_n}{\sqrt{c_n}}\right]$ is stable $(\hat{DW}=1.42, \hat{DF}=6.59, \hat{p}=0$ for ADF), linking hog numbers and the supply of corn.

There are many other possible CI (stable) relationships to be found in this system and while a CI analysis, as outlined in this chapter provides a framework within which to explore such equilibria, we will be mostly concerned with its use to reduce series to stationarity without incurring problems from inducing unit roots, as discussed in §4.2. For this application any three linearly independent CI vectors will be sufficient (e.g. §4.5.1).

4.6 Extensions

There are two extensions to the theory of CI which immediately suggest themselves. The first is when each series requires higher degree differencing to reach stationarity, but some linear combination of them, or of various degrees of differences of each of them is stationary. This was not considered by E&G who restricted attention to the CI(1,1) case. The second extension is when each series requires higher order (e.g. seasonal) differencing to reach stationarity.

4.6.1 Higher degree differencing

Clearly the problem becomes very much larger when we consider higher degree differencing for each series. In addition to having to consider linear combinations of the series, we also have to examine linear combinations of various differences of the series. For instance, suppose series z_{it} requires differencing to degree d_i to make it stationary. We must now consider linear combinations of the form $\alpha^T D_{CI}(B)\mathbf{z}_t$ where α is a $k\times 1$ CI vector and $D_{CI}(B)=\text{diag}(\nabla^{c_i})$, for some $c_i<d_i$, $i=1,\ldots,k$. The number of possible combinations to search over becomes very large for increasing k and d_i, but the general strategy remains the same since

$\mathbf{y}_t = D_{CI}(B)\mathbf{z}_t$ is then another collection of n.s. series with a stationary linear combination α.

As an example of such CI, consider the system

$$\nabla^2 z_{1t} = a_{1t} \iff z_{1t} = 2z_{1,t-1} - z_{1,t-2} + a_{1t}, \qquad [4.22a]$$

$$z_{2t} = \beta z_{1t} + a_{2t}. \qquad [4.22b]$$

z_{1t} and z_{2t} each require second degree differencing to make them stationary (i.e. $d_1 = d_2 = 2$), but $x_t = \alpha^T \mathbf{z}_t = (-\beta, 1)\mathbf{z}_t$ is stationary (in D_{CI}, $c_i = 0$, $i = 1,2$). This system is said to be CI(2,2) – see §4.1. To find such a relationship, we seek a vector α to minimise $\sum x_t^2$ in order to find a linear combination nearest to stationarity (given n.s. components). We then need to test the "equilibrium error" x_t for stationarity. The results of §4.3 will carry through as before and the consistency arguments of E&G will continue to hold (more strongly with I(2) series) since the series satisfy the conditions required. New critical values for the tests will have to be calculated, but we will use those of E&G as guides together with the acf for illustration.

Consider the 3–variable system defined by [4.22a], [4.22b] and $\nabla^2 z_{3t} = a_{3t}$. From a simulation with $\beta = 0.8$ and suitable Σ_a, we can fit the regressions [4.14] with each variable in turn as dependent variable to give the following results.

	vector	$D\hat{W}$	$D\hat{F}$	$A\hat{D}F(p)$	CI?
$\hat{\alpha}^{[1]} =$	$(1, -1.25, 0)^T$	1.99	9.85	$\hat{p}=0$	YES
$\hat{\alpha}^{[2]} =$	$(-0.80, 1, 0)^T$	1.99	9.84	$\hat{p}=0$	YES
$\hat{\alpha}^{[3]} =$	$(-28, 36, 1)^T$ †	0.01	0.15	$0.65, \hat{p}=1$	NO

(† the estimated coefficients in this vector cannot be used since there is no evidence of CI and the estimates are not consistent. However, it confirms that z_{3t} is not included in the CI relationship.) The acfs of $\hat{\alpha}^{[1]T}\mathbf{z}_t$ and $\hat{\alpha}^{[2]T}\mathbf{z}_t$ are both insignificant at all lags and that of $\hat{\alpha}^{[3]T}\mathbf{z}_t$ is slowly decaying, confirming the conclusions. It can be seen that there is only one CI vector ($\hat{\alpha}^{[1]}$ is a multiple of $\hat{\alpha}^{[2]}$) and for this simulation $\hat{\rho}_{12.3} = 1$ and $\hat{\rho}_{23.1} \approx -\hat{\rho}_{31.2} \approx 0.07$, so that the reverse regressions give consistent results. Other simulated CI(2,2) systems which include multiple and more complex CI vectors give similar, very encouraging results.

Another 2–variable system which we might consider is defined by [4.22a] and

$$\nabla z_{2t} = \beta \nabla z_{1t} + a_{2t}. \qquad [4.22b']$$

Here each of the regressions in [4.13] (z_{1t} on z_{2t} and vice versa) fail to find CI (i.e. the residuals are still n.s.), however the regressions

$$\nabla z_{1t} = \alpha_1 \nabla z_{2t} + \varepsilon_{1t} \qquad [4.23a]$$

$$\nabla z_{2t}=\alpha_2 \nabla z_{1t}+\varepsilon_{2t} \qquad [4.23b]$$

find $\hat{\alpha}_1=1.254$ and $\hat{\alpha}_2=0.793$, with both of the residual series stationary (e.g. $D\hat{W}=1.99$). This system is CI(2,1) (since each series requires second degree differencing, but a linear combination of them only requires first), although it is equivalent to considering the usual CI(1,1) situation in terms of the differenced data $\nabla \mathbf{z}_t$ (since each component of $\nabla \mathbf{z}_t$ requires first degree differencing, but a linear combination is stationary), so all of the results and critical values will hold exactly.

It appears that the theory of CI and its analysis can be readily extended to accommodate higher degree differencing in many forms. However, few series in practice will require more than first degree differencing to reduce them to stationarity so that the results for CI(1,1) series will probably be the most useful.

4.6.2 Seasonal CI

It may happen that series each require seasonal differencing, but there is a common seasonal "factor" to their behaviour. Engle et. al. (1989) discuss the combination of short and long–run forecasts of seasonal variables which raises the topic of seasonal CI. They define seasonal CI and consider extending the methods of E&G to deal with this, however the work is still at an early stage and clearly requires a great deal more research (particularly on the testing) before the methods become useful.

Such seasonally CI behaviour may be thought to occur in some situations and its effect on the models will probably be analogous to that of the more usual CI. However, given the complexity introduced when considering seasonality for multivariate series (see §3.1.1), it may be preferable to work with deseasonalised or seasonally adjusted variables where appropriate to avoid difficulties.

4.7 Further topics in CI

In this section we will consider some other topics including some different approaches to dealing with CI among multivariate time series variables.

4.7.1 Forecasting

Engle and Yoo (1987) consider E&G's two stage estimation procedure from the point of view of forecasting – clearly the forecasts from a CI system must be tied together by the CI relationship, so that for example, those from an unrestricted model fitted to the differenced data will diverge, becoming much worse at longer horizons (Engle and Yoo provide simulation results which support this). It is thus essential to impose the constraints implied by the CI relationships onto the

forecasts from the model.

4.7.2 Further test results

Engle and Yoo (1987) also perform further simulations on the test statistics of E&G which confirm that the DW test is very sensitive to the **order** of the null hypothesis (see §4.4.1) and the sample size. The ADF test appears to be robust to such changes and some more critical values are provided (see Table 2 of Engle and Yoo (1987)).

4.7.3 Other approaches to CI

Several other approaches to dealing with CI are presented in the special issue of the Journal of Economic Dynamics and Control (1988) and we briefly discuss some of these below.

Phillips and Ouliaris

Phillips and Ouliaris (1988) criticise E&G's testing procedure since the tests are non–standard and the parameters in the tests are unidentified under the null hypothesis of no CI. In the presence of CI(1,1), the CI vector(s) reduce the variance in the original series by a factor from I(1) to I(0) – it is this reduction which their alternative method seeks to detect. They form the spectral density matrix of the differenced series and use many standard results from multivariate analysis to test for small eigenvalues. Distributional results can be used to obtain critical values and to statistically test the null hypothesis of no CI. A refinement scales the spectral density matrix to be invariant to the units of measurement of the component series (c.f. the correlation matrix) which gives simpler results. However, the (bounds) tests become poor (i.e. wrongly choose CI) when a measure of long–run multiple correlation between the series increases towards 1 (which may often be the case).

Johansen

Johansen (1988) chooses a particular model (VAR(p)) with which to model the CI system, the parameters of which can then be estimated by conditional maximum likelihood. Distributional results are provided and the likelihood ratio test can be applied to find the CI rank r, but we need to choose the order p and are also assuming that a VAR model is a good approximation to the system.

Bossaerts

Bossaerts (1988) believes that E&G's procedure cannot reveal multiple (linearly independent) CI vectors (however see §4.3.3 and §4.3.4), and also criticises the

tests of E&G because the parameters involved are unidentified under the null hypothesis of no CI.

To overcome these difficulties, Bossaerts proposes a new method based upon the canonical correlation analysis between $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$ (see e.g. §5.2.8). In this case we seek vectors $\mathbf{a}$ and $\mathbf{b}$ which maximise

$$\frac{\mathbf{a}^T\Gamma_1^T\mathbf{b}}{((\mathbf{a}^T\Gamma_0\mathbf{a})(\mathbf{b}^T\Gamma_0\mathbf{b}))^{\frac{1}{2}}}$$

(where $\mathbf{a}^T\mathbf{z}_t$ and $\mathbf{b}^T\mathbf{z}_{t-1}$ are the canonical variates and $\Gamma_h=\Gamma(h)$, the covariance matrix of §2.2.1), subject to the constraints $\mathbf{a}^T\Gamma_0\mathbf{a}=\mathbf{b}^T\Gamma_0\mathbf{b}=1$. However, since $\mathbf{z}_t$ is n.s., Γ_0 does not exist asymptotically. Tiao and Tsay (1989) and Bossaerts (1988) show how the problem can be restated in a regression form so that the necessary matrices exist and provide consistent estimates of the canonical vectors (the constraints now become $\mathbf{a}^T\mathbf{a}=\mathbf{b}^T\mathbf{b}=1$).

If the CI rank is r, then we can find r linear combinations α_i of $\mathbf{z}_t$ which are stationary and hence the remaining $k-r$ linearly independent combinations (given by e.g. β_i) will be n.s. – taking $\mathbf{a}=\mathbf{b}=\beta_i$ for the n.s. components will (asymptotically) give us the maximum correlation of 1. Hence the first $k-r$ canonical variates should be maximally correlated and will define the n.s. components, while the remaining r will define CI vectors – since these variates will be stationary. Bossaerts chooses to test $\alpha_i^T\mathbf{z}_t$ for stationarity with Phillips' (1987) unit root test (on the coefficient from an AR(1) fit).

In §5.2.12 we discuss how this method of searching for CI is incorporated within the model specification procedure of Tiao and Tsay (1989).

4.7.4 Conclusions

§4.2 demonstrated the effects which CI has on the VARMA models which can be built for a given set of data and it appears to be preferable to find any CI among the variables (using the techniques of E&G discussed in §4.3 and §4.4) and correct for it in some way (using the transformation and differencing of §4.2.3). §4.5 illustrated the procedure with a case study and §6.4 compares the models which are built for the raw and "corrected" data. Some other approaches have been briefly discussed and clearly a future research project is required to make some comparisons between the competing methods to decide which is preferable. The two–stage procedure of E&G is both straightforward to use and has been shown to give consistent estimates of the CI vectors. It is easy to determine the CI rank in the case of multiple CI vectors and also to extend the procedure to include higher degree differencing as suggested in §4.6.1 (although the extension to

seasonality will inevitably be much more complex).

The important conclusion from this chapter is to recommend that n.s. multivariate time series are tested for CI between the series for three reasons: (1) the relationships may prove interesting themselves (see §4.5.2 and §4.5.3); (2) models for the "corrected" data should not be prone to problems with estimating unit roots – see §4.2; and (3) the forecasts can be constrained to lie "together" in an appropriate way which should improve the forecasting performance of CI models.

Chapter 5. VARMA model specification

5.1 Introduction

In §3.2 we discussed the problems of specifying a VARMA model and a particular concern (§3.2.2) was that of restricting the number of parameters to be estimated in such large models – i.e. specifying a parsimonious model. In the univariate case, order identification methods (see §3.2.1) can be used to suggest which parameters to include in a model (the order), but with multivariate models we can gain considerable benefit from exploiting any parameter redundancy inherent in the matrix coefficients of a VARMA structure (see §5.1.1 and §5.2.3). **Dimension reduction** techniques (§3.2.3) can sometimes be used to reduce k (the number of series to consider), if this is thought to be sensible, but often more general restrictions need to be applied to reduce the number of parameters. The parameter redundancy discussed in §5.1.1 can be exploited by taking a linear transformation of the data and the methods we will consider in this chapter all search for this situation. To do this they specify **minimal** orders for components of the system which can be assembled into a model for the whole system. Thus they each also provide identification of an overall order, although the simplification of the component models is perhaps of greatest importance since it will make the resulting models manageable.

In this chapter we will present and assess three closely–related methods of model specification (due to Tiao and Tsay (1989), Tsay (1989a) and Tsay (1989b)), which all use the covariance structure of the data to indicate possible model orders. We explore and clarify some details of the techniques and consider some possible extensions and the relationships between the methods. With this in mind we can make recommendations as to their use (§5.5.6). Some further techniques will be briefly discussed in §5.6.

5.1.1 Parameter redundancy

The redundancy in parameterisation is assumed to be such that some coefficient matrices Φ_i and Θ_i of a possible VARMA model are not of full rank (although the entries are all significantly different from zero). To illustrate this Tiao and Tsay (1989) give the example VARMA(1,1) model

$$(I-\Phi_1^{[o]}B)\mathbf{z}_t=(I-\Theta_1^{[o]}B)\mathbf{a}_t,\ \Phi_1^{[o]}=\begin{bmatrix}3 & -1\\ 6 & -2\end{bmatrix},\ \Theta_1^{[o]}=\begin{bmatrix}-0.5 & 0.5\\ -1 & 1\end{bmatrix} \qquad [5.1]$$

In this case, the transformed process $\mathbf{y}_t=T\mathbf{z}_t$, $T=\begin{bmatrix}2 & -1\\ -1 & 1\end{bmatrix}$ follows the

VARMA(1,1) model

$$(I-\Phi_1^{[T]}B)\mathbf{y}_t=(I-\Theta_1^{[T]}B)\mathbf{b}_t,\ \Phi_1^{[T]}=T\Phi_1^{[O]}T^{-1}=\begin{bmatrix}0 & 0\\ 2 & 1\end{bmatrix},\ \Theta_1^{[T]}=T\Theta_1^{[O]}T^{-1}=\begin{bmatrix}0 & 0\\ 0 & 0.5\end{bmatrix} \quad [5.2]$$

(where $\mathbf{b}_t=T\mathbf{a}_t$) which has fewer parameters, so that building and estimating model [5.2] in terms of $\mathbf{y}_t$ will simplify the estimation procedure. The simplification is made possible by the nature of the matrices $\Phi_1^{[O]}$ and $\Theta_1^{[O]}$ which are each only of rank 1. We can transform model [5.2] back into model [5.1] for the original data since $\Phi_1^{[O]}=T^{-1}\Phi_1^{[T]}T$ etc. The model specification procedures described in this chapter search for such redundancy together with the necessary transformations to exploit it. It is natural to consider minimal models for each component (see §5.2.1) which together result in the parameterisation of [5.2].

5.2 Tiao and Tsay's Scalar Component method

One method recently proposed by Tiao and Tsay (1989) (=T&T) allows us to specify the order of a VARMA model by considering models for the component series (called Scalar Components – see §5.2.1 below). It also allows us to exploit any possible simplification which can be made due to redundant (in the sense of §5.1.1) parameters by working with a transformation of the data. In this section we will outline the method as presented in T&T, together with some clarifications to the theory, extensions and further consideration. The relationship with some other methods (notably those of Tsay (1989a and 1989b), which are discussed in §5.3 and §5.4) will be investigated in §5.5 and we apply the techniques to some example data in Chapter 6 in order to assess the models produced by them.

5.2.1 Scalar Components

The method proposed by T&T is based upon the idea of scalar components (SCs). These are simply the component (univariate) series of a multiple time series, i.e. $z_{1t},\ldots,z_{kt}$ are the k SCs which make up $\mathbf{z}_t$. We could also take a non–zero linear combination $\mathbf{v}_0^T\mathbf{z}_t$, which would be another SC – notice that the original series' components take this form as well, with e.g. $z_{1t}=\mathbf{v}_0^T\mathbf{z}_t$, $\mathbf{v}_0=(1,0,\ldots,0)^T$ – so we define a SC of $\mathbf{z}_t$ by it's vector $\mathbf{v}_0$. A k–variate multiple time series consists of k SCs which are defined by k linearly independent $\mathbf{v}_0$s. These vectors can be assembled into a transformation matrix T to be applied to the series $\mathbf{z}_t$ (in the usual case of Chapter 2 $T=I$, but in this chapter we consider other possible matrices).

A scalar component model (SCM) is a multivariate model for a SC. In particular, if a multiple series $\mathbf{z}_t$ could be modelled by, say, a VARMA(1,1) model

$$\mathbf{z}_t=\Phi\mathbf{z}_{t-1}-\Theta\mathbf{a}_{t-1}+\mathbf{a}_t, \qquad [5.3]$$

then the SCM for the SC z_{1t} would be the corresponding row of [5.3]

$$z_{1t}=\Phi_{11}z_{1,t-1}+\ldots+\Phi_{1k}z_{k,t-1}-\Theta_{11}a_{1,t-1}-\ldots-\Theta_{1k}a_{k,t-1}+a_{1t}$$

$$\text{or}\quad \mathbf{v}_0^T\mathbf{z}_t+\mathbf{v}_1^T\mathbf{z}_{t-1}=\mathbf{v}_0^T\mathbf{a}_t+\mathbf{u}_1^T\mathbf{a}_{t-1}, \qquad [5.4]$$

where in this case $\mathbf{v}_0^T=(1,0,\ldots,0)$, $\mathbf{v}_1^T=-(\Phi_{11},\ldots,\Phi_{1k})$, $\mathbf{u}_1^T=-(\Theta_{11},\ldots,\Theta_{1k})$. Notice that this is not a univariate model since we have terms from the other SCs in $\mathbf{z}_t$, e.g. $z_{k,t-1}$. A k-variate VARMA model can be specified by k SCMs for the SCs. T&T's method turns the problem of specifying a VARMA model into that of specifying k independent SCs which follow "minimal" SCMs. These SCMs can then be assembled into a parsimonious VARMA representation.

In the same way that the order of a VARMA model [2.2] is (p,q), we have a SCM of order (p,q)

$$\mathbf{v}_0^T\mathbf{z}_t+\mathbf{v}_1^T\mathbf{z}_{t-1}+\ldots+\mathbf{v}_p^T\mathbf{z}_{t-p}=\mathbf{v}_0^T\mathbf{a}_t+\mathbf{u}_1^T\mathbf{a}_{t-1}+\ldots+\mathbf{u}_q^T\mathbf{a}_{t-q} \qquad [5.5]$$

($\mathbf{v}_p\neq\mathbf{0}$, $\mathbf{u}_q\neq\mathbf{0}$), where p is the number of past $\mathbf{z}_t$s included in the model and q the number of past $\mathbf{a}_t$s. Model [5.3] consists of k SCMs, each of order (1,1) (i.e. each in the form of [5.4]), which can be combined to give us a VARMA model of order (1,1). Generally, k SCMs of orders (p_i,q_i) $(i=1,..,k)$ can be combined to give us a VARMA model of order $(p=\max(p_i),q=\max(q_i))$. Also, if we can find k SCs with SCMs of orders (p_i,q_i), such that we have $p_i<p$ or $q_i<q$ for some i, then we will be able to specify a VARMA(p,q) model for these SCs with some zero rows in the parameter matrices – i.e. we will have reduced the number of parameters. The following example demonstrates such sparsity. T&T present a way of finding k SCs and their corresponding SCM orders together with some of the necessary vectors ($\mathbf{v}_0$ etc.), which firstly identifies an overall VARMA model order and can then lead to a simplified estimation problem.

Example

Suppose we have a bivariate series $\mathbf{z}_t$ and we can find vectors to define two SCs which follow SCMs of orders (1,0) and (0,1), say. i.e. we find $\mathbf{v}_0$, $\mathbf{w}_0$ such that

$$\mathbf{v}_0^T\mathbf{z}_t+\mathbf{v}_1^T\mathbf{z}_{t-1}=\mathbf{v}_0^T\mathbf{a}_t$$

$$\mathbf{w}_0^T\mathbf{z}_t=\mathbf{w}_0^T\mathbf{a}_t+\mathbf{w}_1^T\mathbf{a}_{t-1}$$

(for some $\mathbf{v}_1$, $\mathbf{w}_1$). These SCMs can be combined to give us a VARMA(1,1) model in terms of our original data

Model [S]: $$(T-\Phi\lfloor^{s}B)\mathbf{z}_t=(T-\Theta\lfloor^{s}B)\mathbf{a}_t, \qquad [5.6]$$

$$\text{where } T=\begin{bmatrix}\mathbf{v}_0^T\\ \mathbf{w}_0^T\end{bmatrix},\ \Phi_1^{[S]}=\begin{bmatrix}-\mathbf{v}_1^T\\ \mathbf{0}\end{bmatrix},\ \Theta_1^{[S]}=\begin{bmatrix}\mathbf{0}\\ -\mathbf{w}_1^T\end{bmatrix}.$$

Application of the method of T&T will find the vectors $\mathbf{v}_0$ and $\mathbf{w}_0$ as well as providing an initial estimate of $\mathbf{v}_1$, but we will have to estimate the remaining parameters. In this case [5.6] is not necessarily a VARMA model in the usual form, since we are not restricting T to be equal to I as it was in model [2.2]. We can obtain a VARMA model in terms of the SCs (the components of the transformed data $\mathbf{y}_t=T\mathbf{z}_t$) by factorising out the matrix T (T is invertible since the SC vectors are linearly independent)

Model [T]: $$(I-\Phi_1^{[T]}B)\mathbf{y}_t=(I-\Theta_1^{[T]}B)\mathbf{b}_t, \qquad [5.7]$$

where $\Phi_1^{[T]}=\Phi_1^{[S]}T^{-1}$, $\Theta_1^{[T]}=\Theta_1^{[S]}T^{-1}$ and $\mathbf{b}_t=T\mathbf{a}_t$. Notice that we now have a VARMA(1,1) model in the usual form and both of the parameter matrices $\Phi_1^{[T]}$ and $\Theta_1^{[T]}$ will have a zero row (since both $\Phi_1^{[S]}$ and $\Theta_1^{[S]}$ in [5.6] do, and we have only postmultiplied by T^{-1}). However, if we want a VARMA model in terms of our original data (rather than the transformed SCs) we will have to premultiply [5.6] by T^{-1} (still giving us a VARMA(1,1) model)

Model [O]: $$(I-\Phi_1^{[O]}B)\mathbf{z}_t=(I-\Theta_1^{[O]}B)\mathbf{a}_t, \qquad [5.8]$$

where $\Phi_1^{[O]}=T^{-1}\Phi_1^{[S]}=T^{-1}\Phi_1^{[T]}T$ and $\Theta_1^{[O]}=T^{-1}\Theta_1^{[S]}=T^{-1}\Theta_1^{[T]}T$. Now, the parameter matrix $\Phi_1^{[O]}$ does not necessarily contain any zero rows, since we have premultiplied $\Phi_1^{[S]}$ by T^{-1} and so we may have lost the parameter reduction which the method produced. However, estimation of [5.7] in terms of $\mathbf{y}_t$ should give better results than estimation of [5.8], the matrices $X_1^{[O]}$ of which are not of full rank and will give rise to an ill-conditioning problem. Building the model in terms of our SCs, $\mathbf{y}_t$, and transforming back in terms of our original data will be a useful aid to estimation, even if we do not cut down on the number of coefficients in our final model.

5.2.2 Definition of SCMs

We will now look at the techniques necessary to search for "minimal" SCMs and to specify the relevant vectors which define the SCs. The following definition from T&T is needed to enable us to develop the theory of finding SCMs.

Definition 1

A linear combination $y_t=\mathbf{v}_0^T\mathbf{z}_t$ is said to follow an SCM(p,q) if we can find vectors $\mathbf{v}_1,\ldots,\mathbf{v}_p$ such that $u_t=\mathbf{v}_0^T\mathbf{z}_t+\sum_{i=1}^{p}\mathbf{v}_i^T\mathbf{z}_{t-i}$ satisfies

$$E[\mathbf{a}_{t-j}u_t]\begin{cases}\neq\mathbf{0} & j=q\\ =\mathbf{0} & j>q\end{cases} \qquad [5.9]$$

A definition in this form is needed so that we can use the covariance matrices to search for SCMs (see also equation [5.13]). We will show that this is equivalent to the definition [5.5] of §5.2.1, although T&T omit the proof. Throughout we assume that $\mathbf{z}_t$ is stationary with zero mean (or has been made so from differencing/CI analysis (see §4.2.3); T&T show that their theory holds for n.s. series), hence it has MA(∞) representation (§2.1.1)

$$\mathbf{z}_t=\sum_{m=0}^{\infty}\Psi_m\mathbf{a}_{t-m} \qquad [5.10]$$

To find a similar representation for u_t we can write

$$u_t=\sum_{i=0}^{p}\mathbf{v}_i^T\sum_{m=0}^{\infty}\Psi_m\mathbf{a}_{t-m-i}=\sum_{i=0}^{p}\sum_{m=0}^{\infty}\mathbf{v}_i^T\Psi_m\mathbf{a}_{t-m-i},$$

putting $k=m+i$ and swapping the order of the summations we obtain,

$$u_t=\sum_{i=0}^{p}\sum_{k=i}^{\infty}\mathbf{v}_i^T\Psi_{k-i}\mathbf{a}_{t-k}=\sum_{k=0}^{\infty}\left\{\sum_{i=0}^{k}\mathbf{v}_i^T\Psi_{k-i}\right\}\mathbf{a}_{t-k}\quad(\mathbf{v}_i=\mathbf{0},\ i>p)$$

$$=\sum_{k=0}^{\infty}\mathbf{h}_k^T\mathbf{a}_{t-k}, \qquad [5.11]$$

where $\mathbf{h}_k^T=\sum_{i=0}^{k}\mathbf{v}_i^T\Psi_{k-i}$; $\mathbf{h}_0=\mathbf{v}_0$, since $\Psi_0=I$. i.e. u_t can be written as an infinite sum of past $\mathbf{a}_t$s.

If y_t follows an SCM(p,q) in the form [5.5] then $u_t=\sum_{i=0}^{q}\mathbf{u}_i^T\mathbf{a}_{t-i}$ ($\mathbf{u}_0=\mathbf{v}_0$) and

$$E[\mathbf{a}_{t-j}u_t]=E[\mathbf{a}_{t-j}\sum_{i=0}^{q}\mathbf{a}_{t-i}^T\mathbf{u}_i]=\begin{cases}\Sigma_a\mathbf{u}_j & j\le q\\ \mathbf{0} & j>q\end{cases}\quad(\text{since } E[\mathbf{a}_{t-j}\mathbf{a}_{t-i}^T]=\Sigma_a\delta_{ji})$$

(where δ_{ji} is the Kronecker delta and Σ_a is the covariance matrix of the noise $\mathbf{a}_t$). This satisfies [5.9] since Σ_a is positive definite and $\mathbf{u}_q\neq\mathbf{0}$.

If u_t satisfies [5.9] then

$$E[\mathbf{a}_{t-j}u_t]=E\left[\mathbf{a}_{t-j}\left[\sum_{k=0}^{\infty}\mathbf{h}_k^T\mathbf{a}_{t-k}\right]\right]=E\left[\mathbf{a}_{t-j}\left[\sum_{k=0}^{\infty}\mathbf{a}_{t-k}^T\mathbf{h}_k\right]\right]=\Sigma_a\mathbf{h}_j$$

Hence

$$\Sigma_a\mathbf{h}_j\begin{cases}\neq\mathbf{0} & j=q\\ =\mathbf{0} & j>q\end{cases}\Rightarrow\ \mathbf{h}_j\begin{cases}\neq\mathbf{0} & j=q\\ =\mathbf{0} & j>q\end{cases}$$

(since Σ_a is positive definite) which truncates the sum [5.11] to q terms and so

$y_t=\mathbf{v}_0^T\mathbf{z}_t$ follows an SCM(p,q), with

$$u_t=\sum_{i=0}^{p}\mathbf{v}_i^T\mathbf{z}_{t-i}=\sum_{k=0}^{q}\mathbf{h}_k^T\mathbf{a}_{t-k}. \qquad [5.12]$$

□

This also gives us the result that the MA terms in the SCM, $\mathbf{u}_k$, are given by $\mathbf{h}_k$ defined in [5.11]. Definition 1 in the form of [5.9] will prove to be of use, but in order to use the covariance matrix structure to detect SCMs (see later) we need to rewrite this definition in the form

$$E[\mathbf{z}_{t-j}u_t]\begin{cases}\neq\mathbf{0} & j=q\\ =\mathbf{0} & j>q\end{cases} \qquad [5.13]$$

This can also be shown to define an SCM(p,q), but the proof is omitted from T&T. Using the results of [5.10] and [5.11], if y_t follows an SCM(p,q), then $u_t=\sum_{i=0}^{q}\mathbf{u}_i^T\mathbf{a}_{t-i}$ and

$$E[\mathbf{z}_{t-j}u_t]=E[\sum_{m=0}^{\infty}\Psi_m\mathbf{a}_{t-j-m}\sum_{i=0}^{q}\mathbf{a}_{t-i}^T\mathbf{u}_i]=\begin{cases}\sum_{i=j}^{q}\Psi_{i-j}\Sigma_a\mathbf{u}_i & j\leq q\\ \mathbf{0} & j>q\end{cases}$$

which satisfies [5.13] since Σ_a is positive definite, $\Psi_0=I$ and $\mathbf{u}_q\neq\mathbf{0}$.

Also, conversely, if u_t satisfies [5.13] then

$$E[\mathbf{z}_{t-j}u_t]=E[\sum_{m=0}^{\infty}\Psi_m\mathbf{a}_{t-j-m}\sum_{k=0}^{\infty}\mathbf{a}_{t-k}^T\mathbf{h}_k]$$

$$=\sum_{k=j}^{\infty}\Psi_{k-j}\Sigma_a\mathbf{h}_k\begin{cases}\neq\mathbf{0} & j=q\\ =\mathbf{0} & j>q\end{cases}$$

which implies that $\mathbf{h}_k=\mathbf{0}$, $k>q$ and $\mathbf{h}_q\neq\mathbf{0}$ and so y_t follows an SCM(p,q) in the form of [5.5].

□

With this definition we can now develop the tools for finding SCMs, but first we will look at some useful properties of SCMs when they are defined by [5.9] or [5.13].

5.2.3 Further properties

T&T demonstrate some basic properties of SCMs which will be useful when we come to search for them (particularly those concerned with exchangeability of representations).

Lemma 1 of T&T

This states that if we have SCs y_{it}, $i=1,2$ which follow SCM(p_i,q_i)s with structures

$$\sum_{k=0}^{p_i}\mathbf{v}_k^{(i)T}\mathbf{z}_{t-k}=\sum_{k=0}^{q_i}\mathbf{u}_k^{(i)T}\mathbf{a}_{t-k},\ i=1,2 \qquad [5.14(i)]$$

then $y_{3t}=\alpha_1 y_{1t}+\alpha_2 y_{2,t-v}$ ($v\geq 0$, $\alpha_1\neq 0$, $\alpha_2\neq 0$) will follow an SCM(p_3,q_3) where $p_3\leq\max(p_1,p_2+v)$ and $q_3\leq\max(q_1,q_2+v)$. To prove this (omitted from T&T) we use the definition [5.9] with $u_{3t}=\alpha_1 u_{1t}+\alpha_2 u_{2,t-v}=\sum_{k=0}^{p_3}\mathbf{v}_k^{(3)T}\mathbf{z}_{t-k}$ ($\mathbf{v}_k^{(3)}=\alpha_1\mathbf{v}_k^{(1)}+\alpha_2\mathbf{v}_{k-v}^{(2)}$, $k=0,\ldots,p_3$, $\mathbf{v}_i^{(1)}=\mathbf{0}$, $i>p_1$; $\mathbf{v}_i^{(2)}=\mathbf{0}$, $i<0$, $i>p_2$). We then have

$$E[\mathbf{a}_{t-j}u_{3t}]=\alpha_1 E[\mathbf{a}_{t-j}u_{1t}]+\alpha_2 E[\mathbf{a}_{t-j}u_{2,t-v}]=\mathbf{0},\ j>\max(q_1,q_2+v)$$

which fixes $q_3\leq\max(q_1,q_2+v)$ and so y_{3t} will follow at most an SCM(p_3,q_3). □

This allows us to build alternative representations of SCMs and demonstrates how they may be composed of linear combinations of lower order models, stressing the need to build minimal SCMs.

Lemma 2

Lemma 2 of T&T states that if a SC (y_{1t}) has two alternative structures, then there must exist a different SC (y_{2t}) with a simpler SCM than the known ones. An example of such a situation is given in model [5.2] of §5.1.1, where

$$y_{1t}=b_{1t} \text{ and} \qquad [5.15(1)]$$

$$y_{2t}=b_{2t}+2y_{1,t-1}+y_{2,t-1}-0.5b_{2,t-1}. \qquad [5.15(2)]$$

Now [5.15(1)] allows us to replace the term $y_{1,t-1}$ in [5.15(2)] by $b_{1,t-1}$ and so we have two exchangeable SCM(1,1) structures for y_{2t} (rather than "y_{1t}" in the statement of the lemma), with $\mathbf{v}_1^{(1)T}=(-2,-1)$, $\mathbf{u}_1^{(1)T}=(0,-0.5)$ and $\mathbf{v}_1^{(2)T}=(0,-1)$, $\mathbf{u}_1^{(2)T}=(2,-0.5)$.

The proof of lemma 2 given in T&T is terse, but to prove this lemma we let the known SCMs be

$$y_{1t}=\mathbf{v}_0^T\mathbf{z}_t=\mathbf{v}_0^T\mathbf{a}_t-\sum_{k=1}^{p_i}\mathbf{v}_k^{(i)T}\mathbf{z}_{t-k}+\sum_{k=1}^{q_i}\mathbf{u}_k^{(i)T}\mathbf{a}_{t-k},\ i=1,2. \qquad [5.16(i)]$$

Also let $\boldsymbol{\beta}_k=\mathbf{v}_k^{(1)}-\mathbf{v}_k^{(2)}$, $k=1,\ldots,p_0=\max(p_1,p_2)$ ($\mathbf{v}_k^{(i)}=\mathbf{0}, k>p_i$, $i=1,2$)

and $\boldsymbol{\eta}_k=\mathbf{u}_k^{(1)}-\mathbf{u}_k^{(2)}$, $k=1,\ldots,q_0=\max(q_1,q_2)$ ($\mathbf{u}_k^{(i)}=\mathbf{0}, k>q_i$, $i=1,2$)

so $\boldsymbol{\eta}_k=\mathbf{h}_k^{(1)}-\mathbf{h}_k^{(2)}$ (from [5.12] above)

then $\boldsymbol{\eta}_k^T=\sum_{j=0}^{k}\mathbf{v}_j^{(1)T}\Psi_{k-j}-\sum_{j=0}^{k}\mathbf{v}_j^{(2)T}\Psi_{k-j}$ from the definition of $\mathbf{h}_k$ in [5.11] above

$$\text{i.e. } \eta_k^T=\sum_{j=0}^{k}(\mathbf{v}_j^{(1)T}-\mathbf{v}_j^{(2)T})\Psi_{k-j}=\sum_{j=0}^{k}\beta_j^T\Psi_{k-j}. \qquad [5.17]$$

We can find a $v\geq 1$ such that $\beta_v \neq \mathbf{0}$ and $\eta_v \neq \mathbf{0}$ (otherwise the two known SCMs would be the same) and $\beta_i=\mathbf{0}$ for $i<v$ (which is certainly true for $v=1$ since they are the same SC). Then, from [5.17] $\eta_v=\beta_v$ (since $\Psi_0=I$ and $\beta_i=\mathbf{0}$ for $i<v$) and, subtracting [5.16(1)] from [5.16(2)] we have

$$\sum_{k=v}^{p_0}\beta_k^T\mathbf{z}_{t-k}=\sum_{k=v}^{q_0}\eta_k^T\mathbf{a}_{t-k}$$

$$\text{or } \beta_v^T\mathbf{z}_t=\beta_v^T\mathbf{a}_t-\sum_{k=1}^{p_0-v}\beta_{k+v}^T\mathbf{z}_{t-k}+\sum_{k=1}^{q_0-v}\eta_{k+v}^T\mathbf{a}_{t-k}$$

So that $y_{2t}=\beta_v^T\mathbf{z}_t$ is our new SC with SCM of order at most $(p_0-v,\ q_0-v)$ where $v\geq 1$.

□

In the example model from [5.2] discussed above in [5.15], we have two exchangeable forms for y_{2t} (not "y_{1t}") which give $\beta_1=\eta_1=(-2,0)^T$ and define our SCM(0,0) (y_{1t}) as [5.15(1)].

This lemma will prove to be useful when counting SCMs, since it can be used to tell us how many are due to an exchangeable SCM (see §5.2.6).

Redundant parameters

T&T detail a particular case of parameter redundancy which can be avoided by examining the SCM structure of a VARMA model. If, within a VARMA(p,q) model, we have SCM(p_i,q_i)s such that e.g. $p_2>p_1$ and $q_2>q_1$, then only one of the parameters $[\Phi_s]_{21}$ and $[\Theta_s]_{21}$, $s=1,\ldots,\min(p_2-p_1,q_2-q_1)$ is needed. As an example of this, consider a bivariate VARMA(1,1) model consisting of an SCM(0,0) and an SCM(1,1) (e.g. model [5.2]). We can build a model for the transformed data $\mathbf{y}_t$,

$$\begin{bmatrix}y_{1t}\\y_{2t}\end{bmatrix}=\begin{bmatrix}0&0\\\varphi_{21}&\varphi_{22}\end{bmatrix}\begin{bmatrix}y_{1,t-1}\\y_{2,t-1}\end{bmatrix}+\begin{bmatrix}b_{1t}\\b_{2t}\end{bmatrix}-\begin{bmatrix}0&0\\\theta_{21}&\theta_{22}\end{bmatrix}\begin{bmatrix}b_{1,t-1}\\b_{2,t-1}\end{bmatrix}. \qquad [5.18]$$

Now $y_{1t}=b_{1t}$, but the second row of [5.18] gives us

$$y_{2t}=\varphi_{21}y_{1,t-1}+\varphi_{22}y_{2,t-1}+b_{2t}-\theta_{21}b_{1,t-1}-\theta_{22}b_{2,t-1}$$

$$=\varphi_{22}y_{2,t-1}+b_{2t}-(\theta_{21}-\varphi_{21})b_{1,t-1}-\theta_{22}b_{2,t-1},$$

and so we do not need both of the parameters φ_{21} and θ_{21}. This can be straightforwardly generalised (see T&T, section 3.2) and will be particularly useful when building larger models (for large k, p and q).

5.2.4 Specification of SCMs

In order to use the definition [5.13] to find SCMs we expand u_t and evaluate the expectation in [5.13]. The condition on the vectors $\mathbf{v}_0,\ldots,\mathbf{v}_p$ then becomes

$$(\Gamma_j\mathbf{v}_0+\ldots+\Gamma_{j-p}\mathbf{v}_p) \quad \begin{cases} \neq\mathbf{0} & j=q \\ =\mathbf{0} & j>q \end{cases} \qquad [5.19]$$

where $\Gamma_j=E[\mathbf{z}_t\mathbf{z}_{t+j}^T]$ and so $\Gamma_{-j}=\Gamma_j^T$. This can be made more compact by defining $G(m,j)=[\Gamma_j:\cdots:\Gamma_{j-m}]$ and then, the vector $\mathbf{v}=(\mathbf{v}_0^T,\ldots,\mathbf{v}_p^T)^T$ from an SCM(p,q) satisfies

$$G(p,j)\mathbf{v} \quad \begin{cases} \neq\mathbf{0} & j=q \\ =\mathbf{0} & j>q \end{cases} \qquad [5.20]$$

The technique to search for SCMs of "minimal" orders (minimal in the value of $p+q$, for an SCM(p,q)) now involves constructing $G(p,j)$ for various p and testing condition [5.20] for suitable q – such that $p+q=0,1,2,3,\ldots$ The search path is then over the orders (0,0), (0,1), (1,0), (0,2), (1,1) etc. (see table [5.38] and §5.5.5). For a full VARMA model we will need k SCMs, so we will have to find k vs (with independent first k components – "$\mathbf{v}_0$"s) which satisfy [5.20].

To save us having to check condition [5.20] for **all** $j\geq q$ and to fit the problem into a canonical correlations framework, T&T use the following definition

$$\underset{k(h+1)\times k(m+1)}{\Gamma(m,h,j)} = \begin{bmatrix} \Gamma_{j+1} & \cdots & \Gamma_{j+1-m} \\ \vdots & \ddots & \vdots \\ \Gamma_{j+1+h} & \cdots & \Gamma_{j+1+h-m} \end{bmatrix} = \begin{bmatrix} G(m,j+1) \\ \vdots \\ G(m,j+1+h) \end{bmatrix} \qquad [5.21]$$

($h\geq m$ – so that we can solve the multivariate Yule–Walker equations implied by [5.21]; the function of the variable h will be considered later – §5.2.7). Condition [5.20] can be rewritten as

$$G(p,j+1)\mathbf{v} \quad \begin{cases} \neq\mathbf{0} & j=q-1 \\ =\mathbf{0} & j\geq q \end{cases} \qquad [5.22]$$

If $G(p,j+1)$ satisfies [5.22], then so will $\Gamma(p,h,j)$ (since $G(p,j+1)\mathbf{v}=\mathbf{0}$, $j>q$ and $h>0$). The converse also holds. h is taken to be as large as is necessary to ensure that condition [5.22] can be replaced by

$$\Gamma(p,h,q)\mathbf{v} = \mathbf{0} \qquad [5.23]$$

(i.e. condition [5.23] on Γ is testing condition [5.22] for $j=q,\ldots,q+h$ after which [5.22] is taken to hold). Vectors $\mathbf{v}$ which satisfy [5.23] are called **right vectors** corresponding to **zeroes** of $\Gamma(p,h,q)$.

To find SCMs we now construct $\Gamma(p,h,q)$ for $p+q=0,1,2,3,\ldots$ and suitable h

(see §5.2.7) and search for right vectors **v** which define SCMs. T&T suggest that the search can be broken down into two stages:

(1) Finding an overall VARMA model order which involves counting the number of right vectors of $\Gamma(p,h,q)$ for various p and q.

(2) Exploring the detailed SCM structures which constitute this overall order so as to construct a parsimonious parameterisation.

We will illustrate this procedure in later sections (§5.2.9 and §5.2.10).

5.2.5 Nature of the right vectors of Γ

Theorem 1 of T&T details the properties which the vectors associated with a given SCM(p,q) structure have for various orders of the matrix Γ. To illustrate this theorem and show the forms of Γ and its right vectors we display the following matrices. Suppose we have a vector **v** defining an SCM(p,q). Then

$$\Gamma(p,h,q)= \begin{bmatrix} \Gamma_{q+1} & \cdots & \Gamma_{q+1-p} \\ \vdots & & \vdots \\ \Gamma_{q+c+1} & \cdots & \Gamma_{q+c+1-p} \\ \vdots & & \vdots \\ \Gamma_{q+h+1} & \cdots & \Gamma_{q+h+1-p} \end{bmatrix} \qquad [5.24]$$

has one right vector, **v**, from the SCM(p,q). The outline shows where the matrix $\Gamma(p,h,q)$ directly occurs in the following matrices and so where **v** can be used to give a right vector (augmented with zeroes if necessary). For $b\geq 0, c\geq 0$, we have

$$\Gamma(p,h+b,q+c)= \left[\begin{array}{ccc} \Gamma_{q+c+1} & \cdots & \Gamma_{q+c+1-p} \\ \vdots & & \vdots \\ \Gamma_{q+h+1} & \cdots & \Gamma_{q+h+1-p} \end{array}\right] \\ \begin{bmatrix} \vdots & & \vdots \\ \Gamma_{q+c+h+b+1} & \cdots & \Gamma_{q+c+h+b+1-p} \end{bmatrix} \qquad [5.25]$$

and **v** will also be a right vector of this matrix, since h is taken to be large enough that condition [5.22] holds. For $a\geq 0, b\geq 0$, $\Gamma(p+a,h+b,q)$ has the form

$$\begin{array}{|ccc|ccc}
\Gamma_{q+1} & \cdots & \Gamma_{q+1-p} & \cdots & \Gamma_{q+1-p-a} \\
\vdots & & \vdots & & \vdots \\
\Gamma_{q+h+1} & \cdots & \Gamma_{q+h+1-p} & \cdots & \Gamma_{q+h+1-p-a} \\
\hline
\vdots & & \vdots & & \vdots \\
\Gamma_{q+h+b+1} & \cdots & \Gamma_{q+h+b+1-p} & \cdots & \Gamma_{q+h+b+1-p-a}
\end{array} \qquad [5.26]$$

$\Gamma(p,h,q)$ occurs in the top left corner of this matrix, so the vector $\mathbf{v}$ augmented with zeroes, $(\mathbf{v}^T,\underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{\leftarrow(a)\rightarrow})^T$ will be a right vector of this matrix. No other such augmented vectors (i.e. with zeroes at the start of the vector) constructed from $\mathbf{v}$ will be right vectors, otherwise they could be used to define an SCM with MA order lower than q. Also for $a\geq 0, b\geq 0, c\geq 0$ and $a\geq c$, $\Gamma(p+a,h+b,q+c)$ has the form

$$\begin{array}{|ccc|ccc|ccc}
\Gamma_{q+c+1} & \cdots & \Gamma_{q+1} & \cdots & \Gamma_{q+c+1-p} & \cdots & \Gamma_{q+1-p} & \cdots & \Gamma_{q+c+1-p-a} \\
\vdots & & \vdots & & \vdots & & \vdots & & \\
\Gamma_{q+h+1} & \cdots & & \cdots & \Gamma_{q+h+1-p} & \cdots & & \cdots & \vdots \\
\hline
\vdots & \ddots & \Gamma_{q+h+1} & & \cdots & \ddots & \Gamma_{q+h+1-p} & & \\
\hline
\Gamma_{q+c+h+b+1} & \cdots & & \cdots & & \cdots & & \cdots & \Gamma_{q+c+h+b+1-p-a}
\end{array} \qquad [5.27]$$

with the bottom of the matrix $\Gamma(p,h,q)$ occurring in $c+1$ positions from the top left corner and so $\Gamma(p+a,h+b,q+c)$ $(a\geq c)$ has $c+1$ zeroes derived from the SCM(p,q) with corresponding right vectors $(\mathbf{v}^T,\underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{\leftarrow(a)\rightarrow})^T$,...., $(\underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{\leftarrow(c)\rightarrow},\underbrace{\mathbf{v}^T,\mathbf{0}^T,\ldots,\mathbf{0}^T}_{\leftarrow(a-c)\rightarrow})^T$. Similarly, if $a<c$ then $\Gamma(p+a,h+b,q+c)$ is given by

$$\begin{array}{|ccc|ccc|ccc|}
\Gamma_{q+c+1} & \cdots & \Gamma_{q+c+1-a} & \cdots & \Gamma_{q+c+1-p} & \cdots & \Gamma_{q+c+1-p-a} & & \\
\vdots & & \vdots & & \vdots & & \vdots & & \\
\Gamma_{q+h+1} & \cdots & & \cdots & \Gamma_{q+h+1-p} & \cdots & & & \\
\hline
\vdots & \ddots & \Gamma_{q+h+1} & & \cdots & \ddots & \Gamma_{q+h+1-p} & & \\
\hline
 & & \vdots & & & & \vdots & & \\
\Gamma_{q+c+h+b+1} & \cdots & & \cdots & & \cdots & \Gamma_{q+c+h+b+1-p-a} & &
\end{array} \qquad [5.28]$$

and the bottom of the matrix $\Gamma(p,h,q)$ occurs in $a+1$ positions, so that

$\Gamma(p+a,h+b,q+c)$ $(a<c)$ has $a+1$ zeroes derived from the SCM(p,q) with corresponding right vectors $(\mathbf{v}^T,\underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{\leftarrow(a)\rightarrow})^T$,..., $(\underbrace{\mathbf{0}^T,\ldots,\mathbf{0}^T}_{\leftarrow(a)\rightarrow},\mathbf{v}^T)^T$.

Root table

T&T collect the number of zeroes of $\Gamma(m,h,j)$ into a two–way table (over m and j, for suitable h), called the **root table**. With the above considerations from T&T's theorem 1, we can tabulate the pattern of zeroes for a single SCM. For example, an SCM(0,0) gives rise to the pattern in Table 5.1a and an SCM(1,1) to that in Table 5.1b.

Table 5.1 Root tables for example SCMs.

(a) SCM(0,0)

$m\backslash j$	0	1	2	3	4
0	1	1	1	1	1
1	1	2	2	2	2
2	1	2	3	3	3
3	1	2	3	4	4
4	1	2	3	4	5

(b) SCM(1,1)

$m\backslash j$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	1	1	1
2	0	1	2	2	2
3	0	1	2	3	3
4	0	1	2	3	4

The patterns for several SCMs can be added together (since the right vectors are linearly independent) to form a whole root table for Γ (which we will call $R(m,h,j)$). e.g. SCMs of orders (0,0), (0,1) and (1,1) give the following.

$R(m,h,j)=$

$m\backslash j$	0	1	2	3	4
0	1	2	2	2	2
1	1	4	5	5	5
2	1	4	7	8	8
3	1	4	7	10	11
4	1	4	7	10	13

[5.29]

The entry of 4 zeroes at order (1,1) in this table is accounted for by 2 right vectors due to the SCM(0,0) $(=(\mathbf{v}_{(0,0)}^T,\mathbf{0}^T)^T$ and $(\mathbf{0}^T,\mathbf{v}_{(0,0)}^T)^T$, see [5.27] and [5.28]), 1 right vector from the SCM(0,1) $(=(\mathbf{v}_{(0,1)}^T,\mathbf{0}^T)^T$, see [5.26]) and 1 from the SCM(1,1) $(=\mathbf{v}_{(1,1)})$.

Diagonal increment

T&T suggest examining the **diagonal increment** of the root table which is defined to be

$$D(m,h,j)=\begin{cases} R(m,h,j) & \text{if } m=0 \text{ or } j=0 \\ R(m,h,j)-R(m-1,h,j-1) & \text{if } m,j\geq 1 \end{cases}$$

and for the example root table in [5.29] above we have that

$D(m,h,j)=$

$m\backslash j$	0	1	2	3	4
0	1	2	2	2	2
1	1	3	3	3	3
2	1	3	3	3	3
3	1	3	3	3	3
4	1	3	3	3	3

[5.30]

We will have found an overall VARMA model order (m,j) when $D(m,h,j)=k$ (the number of series, or SCs) as can be seen from [5.28] with $a=c=1$ – each SCM contributes 1 **extra** right vector of $\Gamma(m+1,h,j+1)$.

5.2.6 Exchangeability

T&T discuss the case when a SC $\mathbf{v}_0^T\mathbf{z}_t$ has exchangeable SCM(p_1,q_1) and SCM(p_2,q_2) representations with associated vectors $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ (see the discussion of lemma 2 in §5.2.3 and equation [5.16]). We should be able to apply the results of the preceding section to each representation separately and thus the number of zeroes of $\Gamma(m,h,j)$ should be the sum of the numbers due to each representation. However, T&T's theorem 2 states that the number of zeroes is the **maximum** of those due to each SCM. To show this we use T&T's lemma 3 which shows that there is a linear relationship between the right vectors (see also the discussion of T&T).

Lemma 3

Lemma 3 of T&T is rather complicated, but collapses to a simple corollary of lemma 2 upon examination. We have two exchangeable SCMs of orders (p_1,q_1) and (p_2,q_2) and therefore, by lemma 2, we also have an additional SC with SCM of order (p_3,q_3) where $p_3<p_0=\max(p_1,p_2)$, $q_3<q_0=\max(q_1,q_2)$. Assume without loss of generality that $p_1\geq p_2 \Rightarrow p_0=p_1$ (thus $q_1\leq q_2$) and let $v=p_1-p_3$. Lemma 3 then gives us that

$$\begin{vmatrix} \mathbf{v}_0^{(1)} \\ \vdots \\ \mathbf{v}_{p_1}^{(1)} \end{vmatrix} = \begin{vmatrix} \mathbf{v}_0^{(2)} \\ \vdots \\ \mathbf{v}_{p_2}^{(2)} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{vmatrix} + \delta_0 \begin{vmatrix} \mathbf{v}_0^{(3)} \\ \vdots \\ \mathbf{v}_{p_3}^{(3)} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{vmatrix} + \delta_1 \begin{vmatrix} \mathbf{0} \\ \mathbf{v}_0^{(3)} \\ \vdots \\ \mathbf{v}_{p_3}^{(3)} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{vmatrix} + \ldots + \delta_v \begin{vmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{v}_0^{(3)} \\ \vdots \\ \mathbf{v}_{p_3}^{(3)} \end{vmatrix}$$

Lemma 2 gives us that $(\mathbf{v}_j^{(1)}-\mathbf{v}_j^{(2)})=\mathbf{0}$ for $j<v$, i.e. recursively, $\delta_j=0$ for $j<v$. Hence

$$\begin{vmatrix} \mathbf{v}_0^{(1)} \\ \vdots \\ \mathbf{v}_{p_1}^{(1)} \end{vmatrix} = \begin{vmatrix} \mathbf{v}_0^{(2)} \\ \vdots \\ \mathbf{v}_{p_2}^{(2)} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{vmatrix} + \delta_v \begin{vmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{v}_0^{(3)} \\ \vdots \\ \mathbf{v}_{p_3}^{(3)} \end{vmatrix} \quad \text{or} \quad \mathbf{v}^{(1)} = \begin{vmatrix} \mathbf{v}^{(2)} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{vmatrix} + \delta_v \begin{vmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{v}^{(3)} \end{vmatrix} \qquad [5.31]$$

Number of right vectors of Γ for exchangeable SCMs

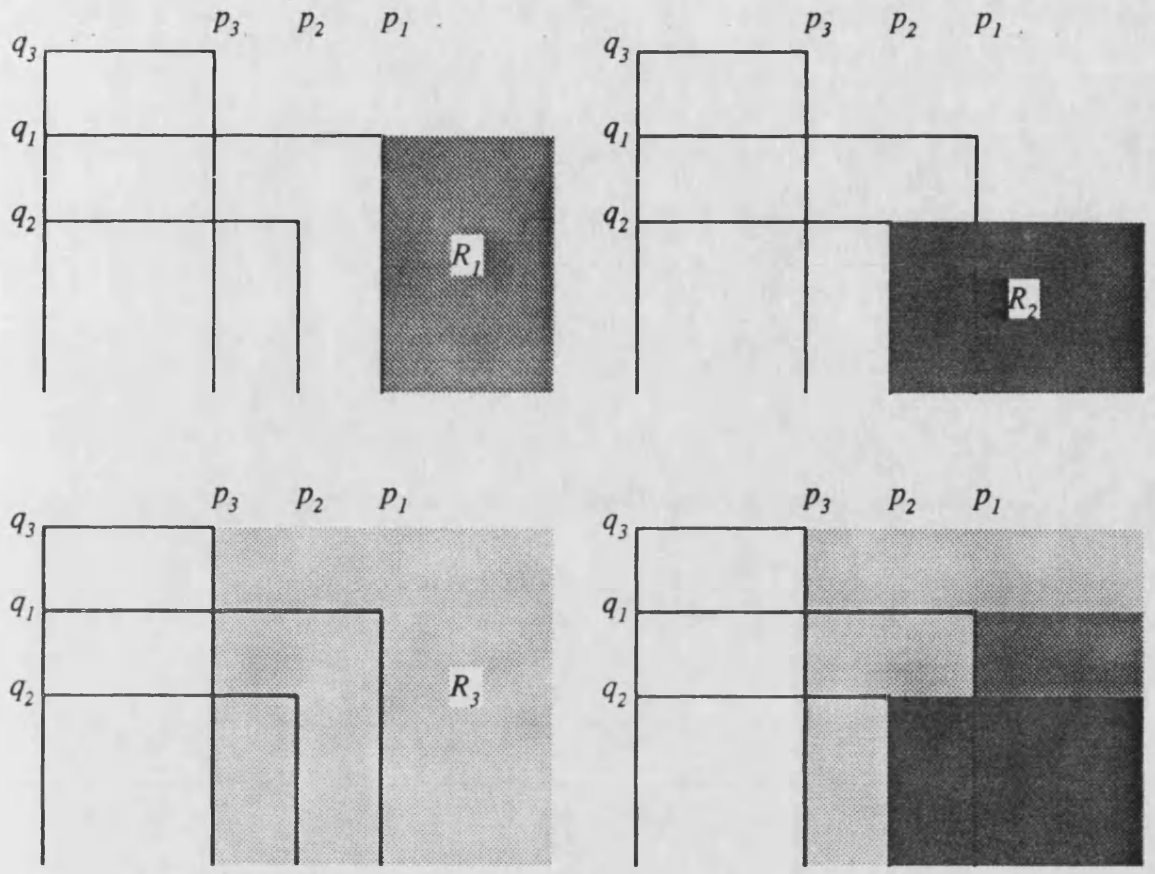
This linear relationship between the vectors restricts the number of linearly independent right vectors of Γ which we can find. To see this consider the constructed covariance matrix $\Gamma(p_3+a,h,q_3+c)$ (for suitably large h), which, for $a=c=0$ has one right vector $\mathbf{v}^{(3)}$. We define $a_i=a-(p_i-p_3)$, $c_i=c-(q_i-q_3)$, $n_i=\min(a_i,c_i)$ and the regions of the root table R_i: $a_i\geq 0$, $c_i\geq 0$. Sketches of these regions within the matrix $\Gamma(p_3+a,h,q_3+c)$ are shown in Figure 5.1a, where the lines indicate the submatrices $\Gamma(p_i,h,q_i)$ (see e.g. [5.27]) and the shaded areas denote the regions. Figure 5.1b also shows the regions in the root table, although here we are displaying the transpose of the usual form ([5.29]). This orientation has the advantage that it corresponds with the matrix Γ if we put Γ_{j+1-m} in the (m,j) position, as can be seen from the form of $\Gamma(m,h,j)$ ([5.21]). If we also define augmented vectors $\mathbf{v}(0,i)=(\mathbf{v}^{(i)T},\underset{\leftarrow(a_i)\rightarrow}{\mathbf{0}^T,\ldots,\mathbf{0}^T})^T,\ldots,$ $\mathbf{v}(n_i,i)=(\underset{\leftarrow(n_i)\rightarrow}{\mathbf{0}^T,\ldots,\mathbf{0}^T},\mathbf{v}^{(i)T},\underset{\leftarrow(a_i-n_i)\rightarrow}{\mathbf{0}^T,\ldots,\mathbf{0}^T})^T$, then we have the pattern of zeroes of $\Gamma(p_3+a,h,q_3+c)$ shown in Table 5.2.

Table 5.2 Number of zeroes.

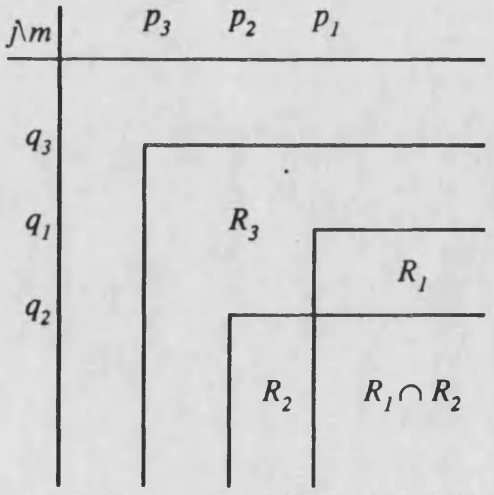
Region	num zeroes	due to	form
$R_3\setminus(R_1\cup R_2)$	n_3+1	$\mathbf{v}^{(3)}$	$\mathbf{v}(0,3),\ldots,\mathbf{v}(n_3,3)$
$R_1\setminus R_2$	n_3+1	$\mathbf{v}^{(3)}$	$\mathbf{v}(0,3),\ldots,\mathbf{v}(n_3,3)$
	$+n_1+1$	$\mathbf{v}^{(1)}$	$\mathbf{v}(0,1),\ldots,\mathbf{v}(n_1,1)$
$R_2\setminus R_1$	n_3+1	$\mathbf{v}^{(3)}$	$\mathbf{v}(0,3),\ldots,\mathbf{v}(n_3,3)$
	$+n_2+1$	$\mathbf{v}^{(2)}$	$\mathbf{v}(0,2),\ldots,\mathbf{v}(n_2,2)$
$R_1\cap R_2$	n_3+1	$\mathbf{v}^{(3)}$	$\mathbf{v}(0,3),\ldots,\mathbf{v}(n_3,3)$
	$+n_2+1$	$\mathbf{v}^{(2)}$	$\mathbf{v}(0,2),\ldots,\mathbf{v}(n_2,2)$
	$+n_1+1$	$\mathbf{v}^{(1)}$	$\mathbf{v}(0,1),\ldots,\mathbf{v}(n_1,1)$

Figure 5.1

(a) Regions of Γ for exchangeable SCMs



(b) Regions of Γ shown in (transposed) root table



However, [5.31] defines linear relationships

$$\mathbf{v}(j,1)=\mathbf{v}(j,2)+\delta_v\mathbf{v}(a_3-a_1+j,3),\quad j=0,..,\min(n_1,n_2),$$

i.e.

$$\begin{bmatrix}\mathbf{0}\\ \mathbf{v}^{(1)}\\ \mathbf{0}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{v}^{(2)}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}+\delta_v\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\\ \mathbf{v}^{(3)}\\ \mathbf{0}\end{bmatrix}$$

which means that $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ between them only contribute $n_1+n_2+1-\min(n_1,n_2)=\max(n_1+1,n_2+1)$ right vectors in the region $R_1\cap R_2$ and hence the number of right vectors for an exchangeable model is the maximum of the numbers for each model at any order.

Example of exchangeable SCMs

Suppose we have a SC with exchangeable SCM(0,1) and SCM(1,0) structures, each with associated vectors $\mathbf{v}_0^{(1)}$ and $\mathbf{v}^{(2)}=(\mathbf{v}_0^{(1)^T},\mathbf{v}_1^{(2)^T})^T$. Lemma 2 then gives us a different SC following an SCM(0,0) with associated vector $\mathbf{v}_0^{(3)}$. We thus have the relations

$$\begin{bmatrix}\Gamma_1\\ \vdots\\ \Gamma_{h+1}\end{bmatrix}\mathbf{v}_0^{(3)}=\mathbf{0},\quad\begin{bmatrix}\Gamma_2\\ \vdots\\ \Gamma_{h+2}\end{bmatrix}\mathbf{v}_0^{(1)}=\mathbf{0}\text{ and }\begin{bmatrix}\Gamma_1 & \Gamma_0\\ \vdots & \vdots\\ \Gamma_{h+1} & \Gamma_h\end{bmatrix}\mathbf{v}^{(2)}=\mathbf{0}.$$

However, when we come to consider the order (1,1), for example

$$\Gamma(1,h,1)=\begin{bmatrix}\Gamma_2 & \Gamma_1\\ \vdots & \vdots\\ \Gamma_{h+2} & \Gamma_{h+1}\end{bmatrix}\text{ has right vectors }\begin{bmatrix}\mathbf{v}_0^{(1)}\\ \mathbf{0}\end{bmatrix},\ \begin{bmatrix}\mathbf{0}\\ \mathbf{v}_0^{(3)}\end{bmatrix}\text{ and }\mathbf{v}^{(2)}=\begin{bmatrix}\mathbf{v}_0^{(1)}\\ \mathbf{v}_1^{(2)}\end{bmatrix}.$$

Lemma 3 gives us a linear relationship between these vectors, specifically $\mathbf{v}_1^{(2)}=\delta_1\mathbf{v}_0^{(3)}$ and so we actually only have two linearly independent right vectors of Γ at order (1,1). Thus, exchangeable SCM(0,1) and SCM(1,0)s will give rise to the root table in Table 5.3a below, and in combination with the SCM(0,0), which must also arise, we have the root table 5.3b.

Table 5.3 Root tables for exchangeable SCMs.

(a) SCM(0,1)+(1,0)

$m\backslash j$	0	1	2	3	4
0	0	1	1	1	1
1	1	1	2	2	2
2	1	2	2	3	3
3	1	2	3	3	4
4	1	2	3	4	4

(b) SCM(0,0)+(0,1)+(1,0)

$m\backslash j$	0	1	2	3	4
0	1	2	2	2	2
1	2	3	4	4	4
2	2	4	5	6	6
3	2	4	6	7	8
4	2	4	6	8	9

5.2.7 Values of h

Theorem 3 of T&T states the results concerning the right vectors of Γ which were illustrated in §5.2.5 above. h is decremented in order to restrict $\Gamma(p+a,h+b,q+c)$ to **only** include the matrix $\Gamma(p,h,q)$, forcing the right vectors to be augmented versions of $\mathbf{v}$. For example, $\Gamma(p,h-c,q+c)$ has the form

$$\Gamma(p,h-c,q+c)=\begin{bmatrix}\Gamma_{q+c+1} & \cdots & \Gamma_{q+c+1-p}\\ \vdots & & \vdots\\ \Gamma_{q+h+1} & \cdots & \Gamma_{q+h+1-p}\end{bmatrix} \qquad [5.32]$$

which is a submatrix of $\Gamma(p,h,q)$ (c.f. [5.24] and [5.25]). Similarly $\Gamma(p+a,h-c,q+c)$ with $a\geq c$ has the form

$$\left[\begin{array}{c|c|c|c}\Gamma_{q+c+1} \cdots & \Gamma_{q+1} \cdots \Gamma_{q+c+1-p} & \cdots \Gamma_{q+1-p} & \cdots \Gamma_{q+c+1-p-a}\\ \vdots & \vdots \quad\quad \vdots & \vdots & \vdots\\ \Gamma_{q+h+1} \cdots & \Gamma_{q+h+1-c} \cdots \Gamma_{q+h+1-p} & \cdots \Gamma_{q+h+1-p-c} & \cdots \Gamma_{q+h+1-p-a}\end{array}\right] \qquad [5.33]$$

(c.f. [5.27]). These results suggest that initially we should take h large so that it can be decremented with each increment in q, which forces the right vectors to be identical. However, the dimension of the matrix Γ increases with h (by multiples of k) and we also want to keep the computational burden down. T&T suggest that when counting right vectors to find an overall VARMA order, $h=m$ will be satisfactory in most cases. An exception to this in when we have a gap in the parameter matrices, which will commonly happen in seasonal cases. T&T give the example of an MA(2) model with $\Theta_1=0$

$$\mathbf{z}_t=\mathbf{a}_t+\Theta_2\mathbf{a}_{t-2},\ \mathbf{a}_t\sim N_k(\mathbf{0},\Sigma_a) \qquad [5.34]$$

(this idea can be extended to seasonal models – see §5.2.11). They call this

phenomenon "skipping lag" and h should be increased (by s – the number of time periods in a season) to cover the gap. To see this, note that $\Gamma_0(=E[\mathbf{z}_t\mathbf{z}_t^T])=\Sigma_a+\Theta_2\Sigma_a\Theta_2^T$, $\Gamma_2=\Sigma_a\Theta_2^T$ and $\Gamma_j=0$, $j=1,3,4,5,\ldots$. Thus $\Gamma(0,0,0)=\Gamma_1=0$ has k independent right vectors, suggesting that the overall VARMA order is (0,0). However, $\Gamma(0,2,j)=\begin{bmatrix}\Gamma_{j+1}\\ \Gamma_{j+2}\\ \Gamma_{j+3}\end{bmatrix}$, so that we would not find k independent right vectors of $\Gamma(0,2,j)$ until the correct order (0,2) was reached (although we may find some lower order SCMs if Θ_2 is not of full rank). Thus, h must be increased whenever skipping lag is suspected, in order to avoid these difficulties.

When we have identified a VARMA model order (P,Q), we will have found k independent right vectors of $\Gamma(P,P,Q)$, so that when seeking SCMs of minimal orders (by examining the detailed SCM structure) we must test condition [5.22] for $j=q,\ldots,Q$ on vectors which are thought to give rise to an SCM(p,q) (since the condition certainly holds for $j>Q$). This (from [5.23]) suggests that h must be at least $Q-q$ and with the constraint that $h\geq p$, in this stage of the analysis we can take $h=\max(p,Q-q)$. T&T suggest taking $h=p+(Q-q)$ $(>\max(p,Q-q))$, but it is not clear what effect these different values will have on the analysis and it may be worth investigating this further.

The problem of finding SCMs now becomes one of constructing $\Gamma(m,h,j)$ for increasing $m+j$ (and suitable h) until we find the diagonal increment, $D(m,h,j)=k$. We then have a VARMA model order and any right vectors of Γ which we find at lower orders will specify parameter reductions in this model for the transformed data (the transformation can be built out of these right vectors – see [5.19] and [5.20], also the example in [5.6]).

5.2.8 Canonical correlations

To make the matrix Γ scale invariant and to facilitate the finding and testing of right vectors, T&T put the problem into a canonical correlations framework. For this we have two sets of variables **D** and **B** and seek linear combinations of them, $\alpha^T\mathbf{D}$ and $\beta^T\mathbf{B}$, say which are maximally correlated – these are our first canonical variates. The next are those linear combinations which are maximally correlated, subject to being uncorrelated with the first, and so on. For the solution to this we seek α,β to maximise

$$\frac{\alpha^T\Sigma_{DB}\beta}{((\alpha^T\Sigma_{DD}\alpha)(\beta^T\Sigma_{BB}\beta))^{\frac{1}{2}}} \qquad [5.35]$$

subject to the constraints $\alpha^T\Sigma_{DD}\alpha=\beta^T\Sigma_{BB}\beta=1$ (see, for example Dillon and

Goldstein (1984, Chapter 9)). The vector α for the first canonical variate can be shown to be an eigenvector of $A=\Sigma_{DD}^{-1}\Sigma_{DB}\Sigma_{BB}^{-1}\Sigma_{BD}$, associated with the largest eigenvalue (where Σ_{BD} is the covariance matrix of **B** and **D**). T&T define $\mathbf{Y}_{m,t}=(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t-m}^T)^T$ and then take $\mathbf{D}=\mathbf{Y}_{m,t}$ and $\mathbf{B}=\mathbf{Y}_{h,t-j-1}$. $\Gamma(m,h,j)$ of [5.21] is then given by $E[\mathbf{Y}_{h,t-j-1}\mathbf{Y}_{m,t}^T]$ (provided that $E[\mathbf{z}_t]=\mathbf{0}$), or $\Gamma(m,h,j)=\Sigma_{BD}$. When $\mathbf{z}_t$ is n.s., Σ_{DD} for example will not exist asymptotically. To overcome this, T&T rewrite the canonical correlations problem in a regression form so that the results can also be shown to hold when $\mathbf{z}_t$ is n.s. Their Theorem 4 states that this form of the problem still gives consistent estimates of the required vectors. A right vector of $\Gamma(m,h,j)$ is also an eigenvector of A associated with a zero eigenvalue. Hence, the problem of finding right vectors of Γ which define SCMs becomes one of finding the αs which give us **zero** canonical correlations between $\mathbf{Y}_{m,t}$ and $\mathbf{Y}_{h,t-j-1}$ and these αs are the vs of [5.20] which we require to define SCMs.

Test Statistics

When working with data, the eigenvalues will not be exactly zero and T&T define a statistic to test whether the s smallest eigenvalues can be taken to be zero. This is related to Bartlett's test of dimensionality (see e.g. Dillon and Goldstein (1984, §9.3.5)) and has the form

$$C(s)= -(N-h-j)\sum_{i=1}^{s}\ln\left[1-\frac{\hat{\lambda}_i}{d_i}\right] \qquad [5.36]$$

where $\hat{\lambda}_i$ is the i'th smallest eigenvalue of $\hat{A}(m,h,j)$, $d_i=1+2\sum_{v=1}^{j}\hat{\rho}_v(\hat{w}_{1t})\hat{\rho}_v(\hat{w}_{2t})$, $\hat{\rho}_v(\hat{w}_{ut})$ is the lag v sample autocorrelation of $\hat{w}_{ut}$ $(u=1,2)$, $\hat{w}_{1t}(i)=\hat{\alpha}_i^T\mathbf{Y}_{m,t}$, $\hat{w}_{2t}(i)=\hat{\beta}_i^T\mathbf{Y}_{h,t}$ (i.e. d_i is a scaling factor for each eigenvalue dependent upon the autocorrelation in the canonical variates (see equation [2.14] of §2.2.1): $\frac{\hat{\lambda}_i}{d_i}$ has unit variance asymptotically). Under the hypothesis that $\lambda_i=0$, $i=1,\ldots,s$, $\lambda_{s+1}\neq 0$ and $w_{1t}(i)$, $i=1,\ldots,s$ is uncorrelated with $\mathbf{a}_{t-v}$, $v>j$ (which is just condition [5.9] for an SCM(m,j) restated using the current notation), T&T's theorem 5 shows that $C(s)$, defined in [5.36] is $\chi^2_{s((h-m)k+s)}$ distributed (Tsay (1989b) provides a complete proof and simulation study comparing the behaviour of $C(s)$ with the usual unscaled statistic). i.e. we reject the hypothesis that $R(m,h,j)\geq s$ if $C(s)>\chi^2_{\alpha,s((h-m)k+s)}$ for some chosen significance level α.

To test whether the diagonal increment $D(m,h,j)$ can be taken to be k, $C(s)$ is altered by T&T to

$$crit(s) = -(N-h-j)\sum_{i=1}^{s+k}\ln\left[1-\frac{\hat{\lambda}_i}{d_i}\right] \qquad [5.37]$$

which is $\chi^2_{(s+k)((h-m)k+s+k)}$ distributed under the hypothesis that $A(m-1,h,j-1)$ has s zero eigenvalues and $\lambda_i=0$, $i=1,\ldots,s+k$, $\lambda_{s+k+1}\neq 0$. We reject the hypothesis that $D(m,h,j)=k$ if $crit(s)>\chi^2_{\alpha,(s+k)((h-m)k+s+k)}$.

5.2.9 Summary of the method of T&T

The method seeks to select suitable orders for (multivariate) models for the k (scalar) components of a multivariate time series. These individual models can then be assembled into a VARMA model for the vector process, in various forms (see §5.2.1). To do this we count candidate vectors which define SCMs of various orders (to construct the root table) and look for patterns in the **increase** in the number of candidate SCM vectors from order (m,j) to $(m+1,j+1)$ (the diagonal increment). We then sort out the detailed SCM structures (i.e. which vectors are due to SCMs of lower orders) in order to specify **sparsity** in the VARMA structure.

Firstly we construct the variables $\mathbf{Y}_{m,t}$, $\mathbf{Y}_{h,t-j-1}$ for $m=0,\ldots,m_0$, $j=0,\ldots,j_0$ (usually $h=m$, $m_0=j_0\approx 4$, but we may choose different values in the presence of seasonality for example) and perform a canonical correlation analysis between them in order to construct the root table $R(m,h,j)$ (using the statistic $C(s)$ of [5.36] and some chosen significance level α). The pattern of diagonal increment $D(m,h,j)$ in this table (judged by the statistic *crit* of [5.37]) will suggest a possible overall VARMA model order, (p,q), say.

We then search for SCMs of minimal orders over $(m+j)=0,1,2,..$ T&T suggest the order $c=0,1,2,..$ $m=0,\ldots,c$, $j=c-m$ which gives us a path over the root table in the order shown below.

$m\backslash j$	0	1	2	3	4
0	1	2	4	7	11
1	3	5	8	12	
2	6	9	13		...
3	10	14			
4	15		...		

[5.38]

We will consider the effects of this search path in §5.5.5. Canonical correlation analysis is also used when each new SCM is found in order to sort out any vectors which duplicate or are due to those already found at lower orders. This ensures that we find linearly independent SCMs.

Models

This analysis gives us k vectors $\mathbf{v}^{(i)}$ of lengths kp_i which are thought to define

the AR parts of the SCM(p_i,q_i)s (i=1,...,k) contained within the overall VARMA(p,q) model. We can assemble these into the matrices

$$T=\begin{bmatrix} \mathbf{v}_0^{(1)T} \\ \vdots \\ \mathbf{v}_0^{(k)T} \end{bmatrix}, \quad \Phi_j^{[s]}= -\begin{bmatrix} \mathbf{v}_j^{(1)T} \\ \vdots \\ \mathbf{v}_j^{(k)T} \end{bmatrix}$$

(where $\mathbf{v}_j^{(i)}$ is a vector consisting of the (jk+1)'st to the (j+1)k'th entries of $\mathbf{v}^{(i)}$ or zeroes if $j>p_i$ – see [5.19] and [5.20]), which give us a model in the form of equation [5.6] of the example in §5.2.1. The entries in the parameter matrices $\Theta_j^{[s]}$ are unknown except for the zero rows specified by the individual q_is (i.e. if $q_i<q$ then row i of $\Theta_j^{[s]}=\mathbf{0}^T$, $j=q_i+1,\ldots,q$). As shown in the example of §5.2.1, there are various forms of the model which we may be interested in. For estimation purposes, we will use the form of equation [5.7] – in terms of the transformed data $T\mathbf{z}_t$, and exploit any sparsity defined by the SCMs. The vectors $\mathbf{v}^{(i)}$ can be used to give us estimates of the AR matrices $\Phi_i^{[T]}$, but we may only wish to use these as starting values for maximum likelihood estimation of the full model. (Bhansali (1989, discussion of T&T) comments that the inverse correlation matrices (e.g. Chatfield (1979)) could be used to obtain estimates of the MA parameters, although the properties of these have not been studied). We can also use the initial estimates of the AR parameters to suggest seasonal constraints – see §5.2.11. The transformed data may themselves be meaningful (defining various relationships between the original variables), but more usually we are interested in a model for the original data $\mathbf{z}_t$. The model in the form of [5.7] can be transformed into one in terms of the original data, e.g. [5.8], but this form may lose the parameter reduction.

5.2.10 Example of the SCM method

In their paper, T&T illustrate the application of their method to the flour price data (Appendix A.2, see also §6.2 for an evaluation of the resulting analysis) and the US hog data (Appendix A.4). Since we have available Tsay's FORTRAN programs for implementing their method, we will also apply it to the annual Lydia Pinkham data (Appendix A.1) for illustration.

Each component of this bivariate dataset is individually non–stationary and the usual tests for CI of Chapter 4 suggest that the components are not CI (see §6.1). We will therefore analyse the differenced series $\nabla\mathbf{z}_t$. The range of ∇z_{1t} (advertising expenditure) is (–712 – 407) and that of ∇z_{2t} (sales) is (–558 – 650) (all in $000) which may be too large for the programs to handle, so we will also normalise the differenced series to each have unit variance and denote this data

by $\tilde{\mathbf{z}}_t$.

Root table (first stage)

Applying the first stage of the method as outlined in preceding sections to $\tilde{\mathbf{z}}_t$, we obtain the following tables of zero eigenvalues of Γ (see [5.29] and [5.30]), judged by the statistics $C(s)$ and $crit(s)$ of [5.36] and [5.37], at 5% significance level.

Root table, $R(m,h,j)$

$m\backslash j$	0	1	2	3	4
0	1	1	2	2	2
1	1	2	3	4	4
2	2	3	4	6	6
3	2	4	5	8	8
4	2	4	6	8	10

Diagonal increment, $D(m,h,j)$

$m\backslash j$	0	1	2	3	4
0	1	1	2	2	2
1	1	1	2	2	2
2	2	2	2	3	2
3	2	2	2	4	2
4	2	2	2	3	2

These results suggest that possible overall VARMA model orders are (0,2) and (2,0) (with SCMs of orders (0,0) and either (0,2) or (2,0)). To see this we give the root tables which would be generated by SCMs of orders (0,0), (0,2), (1,1) and (2,0) in the following tables (we give that for (1,1) to see that the value of 2 at $R(1,h,1)$ cannot be taken to mean that we have found two independent SCMs at order (1,1) – see also §5.2.5).

(0,0)

$m\backslash j$	0	1	2	3	4
0	1	1	1	1	1
1	1	2	2	2	2
2	1	2	3	3	3
3	1	2	3	4	4
4	1	2	3	4	5

(0,2)

$m\backslash j$	0	1	2	3	4
0	0	0	1	1	1
1	0	0	1	2	2
2	0	0	1	2	3
3	0	0	1	2	3
4	0	0	1	2	3

(1,1)

$m\backslash j$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	1	1	1
2	0	1	2	2	2
3	0	1	2	3	3
4	0	1	2	3	4

(2,0)

$m\backslash j$	0	1	2	3	4
0	0	0	0	0	0
1	0	0	0	0	0
2	1	1	1	1	1
3	1	2	2	2	2
4	1	2	3	3	3

These can be added together for any combination of SCMs, so that an SCM(0,0) and a second SCM of order (0,2), (1,1) or (2,0) will give rise to complete root tables as shown below (with the actual one repeated for comparison).

(0,0)+(0,2)

$m\backslash j$	0	1	2	3	4
0	1	1	2	2	2
1	1	2	3	4	4
2	1	2	4	5	6
3	1	2	4	6	7
4	1	2	4	6	8

(0,0)+(1,1)

$m\backslash j$	0	1	2	3	4
0	1	1	1	1	1
1	1	3	3	3	3
2	1	3	5	5	5
3	1	3	5	7	7
4	1	3	5	7	9

(0,0)+(2,0)

$m\backslash j$	0	1	2	3	4
0	1	1	1	1	1
1	1	2	2	2	2
2	2	3	4	4	4
3	2	4	5	6	6
4	2	4	6	7	8

Actual

$m\backslash j$	0	1	2	3	4
0	1	1	2	2	2
1	1	2	3	4	4
2	2	3	4	6	6
3	2	4	5	8	8
4	2	4	6	8	10

However, as discussed in Theorem 2 of T&T and §5.2.5 above, if a SCM can be considered to be either of order (2,0) or (0,2), then each entry in the root table generated by it is obtained by taking the maximum of the corresponding entries in the tables for the exchangeable SCMs. Thus, if we consider exchangeable SCMs (0,2), (2,0) together with the SCM(0,0), we would obtain the following tables.

Root table

$m\backslash j$	0	1	2	3	4
0	1	1	2	2	2
1	1	2	3	4	4
2	2	3	4	5	6
3	2	4	5	6	7
4	2	4	6	7	8

Diagonal increment

$m\backslash j$	0	1	2	3	4
0	1	1	2	2	2
1	1	1	2	2	2
2	2	2	2	2	2
3	2	2	2	2	2
4	2	2	2	2	2

These tables correspond closely with the actual ones obtained for the data $\mathbf{z}_t$, suggesting that the second SCM has exchangeable orders (0,2) and (2,0). We will continue with the analysis, choosing an overall (2,0) order (since e.g. Heyse and Wei (1985b) fitted a VAR(2) model to the differenced data), although we may wish to consider the other possible models and even make a complete analysis of each in order to choose the most appropriate (see §6.1). There does not appear to be an immediate way of choosing between exchangeable models.

SCMs (second stage)

The second stage analysis details the SCM structure for the selected model order, and for m=0,j=0 gives us the following (in the notation of §5.2.8 – the p-values refer to the upper percentage point of the relevant χ^2 distribution)

s	$\hat{\lambda}_s$	$\hat{C}(s)$	dof	$\chi^2_{0.95}$	p-value
1	0.027	1.46	1	3.84	0.23
2	0.195	12.94	4	9.49	0.01

These results suggest that (with a 5% significance level) we have found one right vector of $\Gamma(0,0,0)$, giving us an SCM(0,0). This vector is $\mathbf{v}^{(1)T}=(0.79,-0.62)$. At the order (1,0), we only find one right vector of Γ, which is due to that at the order (0,0) and so we do not have any SCM(1,0)s. At the order (2,0) we obtain the following results

s	$\hat{\lambda}_s$	$\hat{C}(s)$	dof	$\chi^2_{0.95}$	p-value
1	0.0002	0.01	1	3.84	0.91
2	0.109	5.92	4	9.49	0.21

One of the right vectors is due to $\mathbf{v}^{(1)}$, but the other is not linearly dependent on it, so now we have found two SCMs (of orders (0,0) and (2,0)) with the right vector corresponding to the SCM(2,0) given by $\mathbf{v}^{(2)}$. These results suggest that we can build a VAR(2) model for $T\mathbf{z}_t$ (see the example in §5.2.1), where

$$T=\begin{bmatrix}0.79 & -0.62\\ 0.62 & 0.79\end{bmatrix},\ \Phi_1^{[T]}=\begin{bmatrix}0 & 0\\ -0.65 & 0.47\end{bmatrix},\ \Phi_2^{[T]}=\begin{bmatrix}0 & 0\\ -0.45 & -0.31\end{bmatrix}. \qquad [5.39]$$

(The coefficients in the Φ_i matrices can be estimated from the SCM vectors $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ as illustrated in §5.2.1.)

The SCM analysis has enabled us to specify a reduced VAR(2) model which will be easier to estimate than an unrestricted one (see §6.1 and §6.2 for more discussion of this aspect of the analysis).

5.2.11 Modelling seasonality

T&T did not consider seasonal aspects of SCMs. In this section we show how their results can be used to deal with such models.

The method of T&T only specifies the **orders** of k SCMs which we use to build a VARMA model. The order of an SCM is the **maximum** number of past $\mathbf{z}_t$s and $\mathbf{a}_t$s included in the model. However, the structure of an SCM(p,q) in the form of [5.5], may contain vectors $\mathbf{v}_i$, $0<i<p$ or $\mathbf{u}_i$, $0<i<q$ which are zero (a situation referred to in §5.2.7 above as "skipping lag") and hence only specifying overall orders may lead to overparameterisation. Seasonal models may be thought of as instances of skipping lag. For instance, we may have a VARMA(1,0)×(1,0)$_{12}$ model

$$(I-\Phi_1 B)(I-\Phi_{12}B^{12})\mathbf{z}_t=\mathbf{a}_t, \qquad [5.40]$$

with, for example k=2, and

$$\Phi_1=\begin{bmatrix} 0.27 & -0.30 \\ -0.23 & 0.37 \end{bmatrix},\ \Phi_{12}=\begin{bmatrix} 0.76 & 0.0 \\ 0.0 & 0.69 \end{bmatrix},\ \Sigma_a=\begin{bmatrix} 4.2 & \\ 1.6 & 4.1 \end{bmatrix}\times 10^{-3}. \quad [5.41]$$

This needs to be multiplied out to put it into a SCM framework (since SCMs are additive, not multiplicative), which gives a VARMA(13,0) model

$$\mathbf{z}_t=\Phi_1\mathbf{z}_{t-1}+\Phi_{12}\mathbf{z}_{t-12}-\Phi_1\Phi_{12}\mathbf{z}_{t-13}+\mathbf{a}_t. \quad [5.42]$$

This consists of k SCM(13,0)'s

$$\mathbf{v}_0^T\mathbf{z}_t+\mathbf{v}_1^T\mathbf{z}_{t-1}+\mathbf{v}_{12}^T\mathbf{z}_{t-12}+\mathbf{v}_{13}^T\mathbf{z}_{t-13}=\mathbf{v}_0^T\mathbf{a}_t, \quad [5.43]$$

but of the 14 vectors necessary to describe an SCM of order (13,0), in this case 10 are equal to $\mathbf{0}$ ($\mathbf{v}_2,\ldots,\mathbf{v}_{11}$).Specifying only the order of an SCM takes no account of these zero vectors and would result in overparameterisation. i.e. if we identified k SCM(13,0)'s from data which follows model [5.42], then, applying the method as presented in T&T, we would try to estimate all of the matrices Φ_1, Φ_2, Φ_3,..., Φ_{13} and inevitably meet problems.

However, the canonical correlation analysis which is performed in order to identify the SCM orders also gives "estimates" of all the (AR) vectors, $\mathbf{v}_i$, in the SCMs. Hence it may be that the zero vectors ($\mathbf{v}_2,\ldots,\mathbf{v}_{11}$) will be identified as such and the final estimation of a full model can be constrained to the form of [5.42].

For the example model [5.42]+[5.41] above we can calculate the theoretical covariance matrices (see equations [2.17] in §2.2.2), which are all that are needed to perform the canonical correlation analysis (i.e. we have not simulated data from this model, so the analysis is exact, except for the error due to the iterations required to estimate the covariance matrices). This results in two zero eigenvalues of $A(13,13,0)$ (the relevant matrix used in the analysis – see §5.2.8) which correctly identifies the order (notice that using e.g. $h=m$ may have falsely suggested that the SCMs were of lower order). The root table $R(m,13,j)$ is given below.

$m\backslash j$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
⋮		⋮			⋮	
12	0	0	0	0	0	0
13	2	2	2	2	2	2
14	2	4	4	4	4	4
15	2	4	6	6	6	6
16	2	4	6	8	8	8

The eigenvectors associated with the two zeroes take the form

$$\mathbf{v}^{(i)} = \left[\mathbf{v}_0^{(i)T},\mathbf{v}_1^{(i)T},\mathbf{0}^T,\ldots,\mathbf{0}^T,\mathbf{v}_{12}^{(i)T},\mathbf{v}_{13}^{(i)T}\right]^T,\ i=1,2$$

(where the **0** vectors are not identically zero, but none of their components are greater than about 10^{-6} in magnitude). i.e. we have the two SCMs

$$\mathbf{v}_0^{(i)T}\mathbf{z}_t+\mathbf{v}_1^{(i)T}\mathbf{z}_{t-1}+\mathbf{v}_{12}^{(i)T}\mathbf{z}_{t-12}+\mathbf{v}_{13}^{(i)T}\mathbf{z}_{t-13}=\mathbf{v}_0^{(i)T}\mathbf{a}_t. \qquad [5.44(i)]$$

If we now take

$$P_i=\begin{bmatrix}\mathbf{v}_i^{(1)T}\\ \mathbf{v}_i^{(2)T}\end{bmatrix},\ i=0,1,12,13,\ P_0=T \text{ (the transformation)}$$

and combine [5.44(1)] and [5.44(2)], we obtain

$$P_0\mathbf{z}_t=-P_1\mathbf{z}_{t-1}-P_{12}\mathbf{z}_{t-12}-P_{13}\mathbf{z}_{t-13}+P_0\mathbf{a}_t. \qquad [5.45]$$

We can transform this model back (by premultiplying [5.45] by P_0^{-1}) to obtain a VARMA model in terms of $\mathbf{z}_t$

$$\mathbf{z}_t=-P_0^{-1}P_1\mathbf{z}_{t-1}-P_0^{-1}P_{12}\mathbf{z}_{t-12}-P_0^{-1}P_{13}\mathbf{z}_{t-13}+\mathbf{a}_t \qquad [5.46]$$

In this case we find that $-P_0^{-1}P_i=\Phi_i$ (to within about 0.01), $i=1,12,13$, so that T&T's technique has allowed us to correctly identify (and also to "estimate" the coefficient matrices of) model [5.42]+[5.41], including the zero matrices. Obviously this will not happen so well in practice and some investigation is needed to evaluate the effectiveness with real data, however, this example serves to illustrate that perhaps we do not need to **extend** the method of T&T to handle seasonal models, but merely to use the extra information contained in the eigenvectors.

T&T's suggested extension for the seasonal case

T&T's suggestion (in their reply to the discussion) for handling the seasonal case is to extend the definition of the constructed variable $\mathbf{Y}_{m,t}$ used in the analysis to that of

$$\mathbf{Y}_{m_1,m_2,t}=(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t-m_1}^T,\mathbf{x}_{m_1,t-s}^T,\ldots,\mathbf{x}_{m_1,t-m_2s}^T)^T,$$

where $\mathbf{x}_{m_1,t}=(\mathbf{z}_{t+m_1}^T,\ldots,\mathbf{z}_t^T,\ldots,\mathbf{z}_{t-m_1}^T)^T$, so that

$$\mathbf{Y}_{m_1,m_2,t}=(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t-m_1}^T,\mathbf{z}_{t-s+m_1}^T,\ldots,\mathbf{z}_{t-s-m_1}^T,\ldots,\ \mathbf{z}_{t-m_2s+m_1}^T,\ldots,\mathbf{z}_{t-m_2s-m_1}^T)^T$$

m_1=AR order, m_2=AR$_s$ order in a multiplicative VAR model. This now gives us three variables m_1, m_2 and j to consider which makes the extension of the root table of §5.2.5 non–trivial! To see how this suggestion might work, consider the model given in [5.40]. For this model then, $m_1=1$, $m_2=1$, $s=12$, so that to correctly identify it we need to consider the variables

$$\mathbf{D}^{[E]}=\mathbf{Y}_{1,1,t}=(\mathbf{z}_t^T,\mathbf{z}_{t-1}^T,\mathbf{z}_{t-11}^T,\mathbf{z}_{t-12}^T,\mathbf{z}_{t-13}^T)^T \text{ and } \mathbf{B}^{[E]}=\mathbf{Y}_{1,1,t-1}.$$

That is, we would expect to find k zero canonical correlations between $\mathbf{D}^{[E]}$ and

$\mathbf{B}^{[E]}$. i.e. there should be k linearly independent right vectors, $\mathbf{v}$, of $\Sigma_{BD}^{[E]}$, where

$$\Sigma_{BD}^{[E]}=\left[\begin{array}{cc|ccc} \Gamma_1 & \Gamma_0 & \Gamma_{10}^T & \Gamma_{11}^T & \Gamma_{12}^T \\ \Gamma_2 & \Gamma_1 & \Gamma_9^T & \Gamma_{10}^T & \Gamma_{11}^T \\ \hline \Gamma_{12} & \Gamma_{11} & \Gamma_1 & \Gamma_0 & \Gamma_1^T \\ \Gamma_{13} & \Gamma_{12} & \Gamma_2 & \Gamma_1 & \Gamma_0 \\ \Gamma_{14} & \Gamma_{13} & \Gamma_3 & \Gamma_2 & \Gamma_1 \end{array}\right], \mathbf{v}=\begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \mathbf{v}_{11} \\ \mathbf{v}_{12} \\ \mathbf{v}_{13} \end{bmatrix}$$

The original procedure uses variables $\mathbf{Y}_{m,t}$ and $\mathbf{Y}_{h,t-j-1}$ and to identify model [5.40] (in particular, [5.42], when multiplied out) we would consider variables of order $m=13, j=0$ (with $h=13$), i.e.

$$\mathbf{D}^{[O]}=\mathbf{Y}_{13,t}=(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t-13}^T)^T \text{ and } \mathbf{B}^{[O]}=\mathbf{Y}_{13,t-1}$$

The canonical correlation analysis between $\mathbf{D}^{[O]}$ and $\mathbf{B}^{[O]}$ gives right vectors of $\Sigma_{BD}^{[O]}$, where

$$\Sigma_{BD}^{[O]}=\left[\begin{array}{cc|c|cc} \Gamma_1 & \Gamma_0 & & \Gamma_{11}^T & \Gamma_{12}^T \\ \Gamma_2 & \Gamma_1 & \cdots & \Gamma_{10}^T & \Gamma_{11}^T \\ \hline \vdots & & \ddots & \vdots & \\ \hline & & & & \\ \Gamma_{13} & \Gamma_{12} & & \Gamma_1 & \Gamma_0 \\ \Gamma_{14} & \Gamma_{11} & \cdots & \Gamma_2 & \Gamma_1 \\ & & & & \end{array}\right], \mathbf{u}=\begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \mathbf{0} \\ . \\ \mathbf{0} \\ \mathbf{v}_{12} \\ \mathbf{v}_{13} \end{bmatrix}$$

So that a right vector $\mathbf{u}$ of $\Sigma_{BD}^{[O]}$ gives a right vector $\mathbf{v}$ of $\Sigma_{BD}^{[E]}$ (but not necessarily vice versa since e.g. here perhaps $\mathbf{v}_{11}\neq\mathbf{0}$). This suggests that T&T's suggested extension can be accommodated within the usual formulation of the method and is probably unnecessary as a special case. Furthermore, the multiplicative aspect (which accounts for the difference between $[E]$ and $[O]$) does not fit naturally into the SCM framework and the extension makes an already complicated procedure perhaps too complex.

Although we could build seasonal models using T&T's (original) method, the extra complexity introduced (see §2.1.2) might deter us from attempting to and we may prefer to seasonally adjust the series (see §3.1.1(d)).

5.2.12 Co–integration

Chapter 4 considered CI between n.s. variables and in §4.7.3 we briefly described Bossaerts' (1988) method which uses the canonical correlation analysis between $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$ to search for CI. In the method of T&T, the analysis at

order $(m,h,j)=(0,0,0)$ uses exactly the same variables and can thus be used to determine CI relationships, if they exist.

Examples

(i) Example (b) of §4.3.1

With data simulated from this example, T&T's procedure identifies one SCM(0,0) (with a p–value of 0.856, against a chosen significance level of 0.05). This has associated with it a vector $\mathbf{v}_0^{(1)T}= (-0.59,0.81,-0.01)= 0.81(-0.73,1,-0.02) \approx 0.81\alpha$ of example (b) of §4.3.1. We will test the transformed series for stationarity and hence for the existence of CI, together with those from the other examples at the end of this section.

(ii) Example (c) of §4.3.1

Using this example now, T&T identifies one SCM(0,0) (p=0.856) with $\mathbf{v}_0^{(2)T}=(-0.54,0.73,0.43) =0.73(-0.73,1,0.59) \approx 0.73\alpha$ of this example.

(iii) The example of §4.3.4

T&T identifies two SCM(0,0)s (p=0.276,0.237) with $\mathbf{v}_0^{(3)T}= (-0.24,0.78,0.58)\approx -0.37\alpha_1+0.97\alpha_2$ and $\mathbf{v}_0^{(4)T}=(0.68,-0.43,0.59)\approx -1.58\alpha_1+0.99\alpha_2$ (α_i vectors from the example).

Now we test the series $\mathbf{v}_0^{(i)T}\mathbf{z}_t$, for stationarity using the tests presented in §4.4.1.

Vector	$\hat{DW}$	$\hat{DF}$	$ADF(\hat{p})$	Stationary series?
$\mathbf{v}_0^{(1)}$	1.44	7.47	$\hat{p}=0$	YES
$\mathbf{v}_0^{(2)}$	1.44	7.48	$\hat{p}=0$	YES
$\mathbf{v}_0^{(3)}$	1.25	6.65	$\hat{p}=0$	YES
$\mathbf{v}_0^{(4)}$	1.68	11.76	$\hat{p}=0$	YES

This demonstrates that T&T can find CI vectors if they exist, although, as discussed in §4.7.4 we prefer to use the regression approach of Engle and Granger (1987) which permits refinement of the CI vectors.

5.3 Tsay's (1989a) Kronecker Index approach

Tsay (1989a) (=T89a) outlines another approach to the specification of VARMA models based upon what are called the Kronecker Indices of the component series (see §5.3.1). T89a demonstrates the close connection of this method with that of T&T, although it does not have the same flexibility over the choice of model order as T&T. In this section we will consider some details of the Kronecker Index approach and it's relationship with the method of T&T. We

will also explore this further in §5.5.

5.3.1 Kronecker Indices

T89a defines ψ_{t-j} to be the σ–field generated by $\{\mathbf{a}_{t-j},\mathbf{a}_{t-j-1},\dots\}$, so that ψ_{t-j} contains all of the information within this set, i.e. if x_t is a random variable, then

$$E[x_t \mid \psi_{t-j}]=E[x_t \mid \{\mathbf{a}_{t-j},\mathbf{a}_{t-j-1},\dots\}].$$

T89a then defines the i'th Kronecker Index (KI) of the vector process $\mathbf{z}_t$ to be the minimum non–negative integer ν, such that the forecast

$$z_{i,t+\nu|t-1}=E[z_{i,t+\nu} \mid \psi_{t-1}]$$

is a linear combination of the previous forecasts $\{z_{i-1,t+\nu|t-1},\dots,z_{1,t+\nu|t-1}, \mathbf{z}_{t+\nu-1|t-1},\dots,\mathbf{z}_{t|t-1}\}$, i.e. there exist vectors $\mathbf{v}_i$, such that

$$E[\sum_{i=0}^{\nu}\mathbf{v}_i^T\mathbf{z}_{t+\nu-i} \mid \psi_{t-1}]=\sum_{j=0}^{\nu}\mathbf{v}_j^T\mathbf{z}_{t+\nu-j|t-1}=0 \qquad [5.47]$$

(where $\mathbf{v}_0^T=(\nu_{01},\nu_{02},\dots,\underset{(i)}{1},0,\dots,0)$). This is the usual definition of KIs (see for example Solo (1986)), which are also called output indices or output lags and are used in the specification of linear systems (see for example Hannan and Diestler (1988)). The KIs measure the duration of the components' dependencies on the past and are not necessarily associated with a particular component series as an example below will demonstrate. The sum of the KIs is called the **McMillan Degree**. To find the KIs of a vector process, we examine the row dependencies of the constructed covariance matrix

$$\underset{kh\times kh}{H_h}=\begin{bmatrix}\Gamma_1^T & \Gamma_2^T & \cdots & \Gamma_h^T\\ \Gamma_2^T & \Gamma_3^T & \cdots & \Gamma_{h+1}^T\\ \vdots & \vdots & \ddots & \vdots\\ \Gamma_h^T & \Gamma_{h+1}^T & \cdots & \Gamma_{2h-1}^T\end{bmatrix}=E[\mathbf{F}_{h-1,t}\mathbf{Y}_{h-1,t-1}^T] \qquad [5.48]$$

(where $\mathbf{F}_{m,t}=(\mathbf{z}_t^T,\dots,\mathbf{z}_{t+m}^T)^T$ is the **future** vector, $\mathbf{Y}_{m,t}$ is the **past** vector as defined in §5.2.8 and h is taken to be larger than an anticipated maximum ν). Row $kn+i$ of H_h is given by

$$E[z_{i,t+n}\mathbf{Y}_{h-1,t-1}^T].$$

If this row is linearly dependent on the previous rows, then there will exist vectors $\mathbf{v}_j$ such that

$$E[\sum_{j=0}^{n}\mathbf{v}_j^T\mathbf{z}_{t+n-j}\mathbf{Y}_{h-1,t-1}^T]=0 \qquad [5.49]$$

($\mathbf{v}_0$ as in [5.47]) and so

$$E[E[\sum_{j=0}^{n}\mathbf{v}_j^T\mathbf{z}_{t+n-j}\mathbf{Y}_{h-1,t-1}^T\,|\,\psi_{t-1}]]=E[\sum_{j=0}^{n}\mathbf{v}_j^T\mathbf{z}_{t+n-j|t-1}\mathbf{Y}_{h-1,t-1}^T]=0$$

(since $\mathbf{Y}_{h-1,t-1}$ will be determined by ψ_{t-1}), which implies that [5.47] holds. Conversely, row $kn+i$ can also be written as

$$E[E[z_{i,t+n}\mathbf{Y}_{h-1,t-1}^T\,|\,\psi_{t-1}]]=E[z_{i,t+n|t-1}\mathbf{Y}_{h-1,t-1}^T] \qquad [5.50]$$

and if [5.47] holds, then we can find vectors $\mathbf{v}_j$ such that row $kn+i$ is linearly dependent on the previous rows. Hence the row dependencies in H_h will define the KIs – v_i is the minimum n such that row $kn+i$ is linearly dependent on the previous rows. We thus examine them from the top down. Note that if row u is linearly dependent on the previous ones, then rows $kr+u$ ($r\geq 1$) will also be (from [5.49], since $\mathbf{z}_t$ is stationary) and we can exclude them from consideration – this automatically eliminates any redundancy such as is inherent in the method of T&T (see §5.2.9 and §5.5.4).

Example

As an example of the row dependencies in H_h (h=5, taken from T89a) and the resulting KIs consider the following matrix for a 5–variate process (with the rows grouped together in columns of 5),

Row	Dep	Row	Dep	Row	Dep	Row	Dep	Row	Dep
1	X	6	X	11	0	16	0	21	0
2	0	7	0	12	0	17	0	22	0
3	X	8	0	13	0	18	0	23	0
4	X	9	X	14	X	19	X	24	0
5	X	10	X	15	0	20	0	25	0

where "X" is used to indicate that this row is independent of it's predecessors and "0" that it is dependent on them. We can transpose these dependencies to give the following.

		component				
		1	2	3	4	5
	0	X	0	X	X	X
	1	X	0	0	X	X
"lag"	2	0	0	0	X	0
	3	0	0	0	X	0
	4	0	0	0	0	0

(=D) [5.51]

(i.e. $D(j,i)$ indicates the dependency of row $kj+i$ on the previous rows of H_h.) From this (with the previous considerations) the KIs can be read off as the number of lags for which we have dependent rows ("X"s) for each component (column) – i.e. $\{v_i\}_{i=1,\ldots,5}=\{2,0,1,4,2\}$.

AR parameters

Consider the usual VARMA representation of [2.2]

$$\mathbf{z}_t-\Phi_1\mathbf{z}_{t-1}-\ldots-\Phi_p\mathbf{z}_{t-p}=\mathbf{a}_t-\Theta_1\mathbf{a}_{t-1}-\ldots-\Theta_q\mathbf{a}_{t-q}.$$

By stationarity, for $t+s$, $s\geq\max(p,q)$ this can also be written as

$$\Xi_s\mathbf{F}_{s,t}=\Upsilon_s\mathbf{B}_{s,t}, \qquad [5.52]$$

where $\Xi_s=\left[\underset{\leftarrow(s-p)\rightarrow}{0:\cdots:0}:-\Phi_p:-\Phi_{p-1}:\cdots:I\right]$, $\mathbf{F}_{s,t}=\left(\mathbf{z}_t^T,\ldots,\mathbf{z}_{t+s}^T\right)^T$,

$\Upsilon_s=\left[\underset{\leftarrow(s-q)\rightarrow}{0:\cdots:0}:-\Theta_q:-\Theta_{q-1}:\cdots:I\right]$, $\mathbf{B}_{s,t}=\left(\mathbf{a}_t^T,\ldots,\mathbf{a}_{t+s}^T\right)^T$.

If we postmultiply [5.52] by $\mathbf{Y}_{h-1,t-1}^T$ and take expectations we obtain

$$\Xi_s\begin{bmatrix}\Gamma_1^T & \cdots & \Gamma_h^T\\ \vdots & \ddots & \vdots\\ \Gamma_{s+1}^T & \cdots & \Gamma_{h+s}^T\end{bmatrix}=E[\Upsilon_s\mathbf{B}_{s,t}\mathbf{Y}_{h-1,t-1}^T]=0 \qquad [5.53]$$

(since the newest element in $\mathbf{Y}_{h-1,t-1}$ is $\mathbf{z}_{t-1}$ and the oldest in $\mathbf{B}_{s,t}$ is $\mathbf{a}_t$) where the constructed covariance matrix is a submatrix (the first $s+1$ block rows) of H_h (if $s<h$). This suggests that a row of Ξ_s gives us a linear combination of the rows of H_h (up to block row $s+1$ – i.e. row $k(s+1)$) which is zero. Alternatively, the row linear dependencies in H_h can be used to indicate which autoregressive parameters are required in a VARMA model for the data (although, now we allow a contemporaneous coefficient matrix Φ_0). A dependent row (kv_i+i) gives a linear combination $\mathbf{V}_s^{(i)T}\mathbf{F}_{s,t}$ $\left(=\sum_{j=0}^{v_i}\mathbf{v}_j^{(i)T}\mathbf{z}_{t+s-j},\ \mathbf{V}_s^{(i)}=(\underset{\leftarrow(s-v_i)\rightarrow}{\mathbf{0}^T,\ldots,\mathbf{0}^T},\mathbf{v}_{v_i}^{(i)T},\ldots,\mathbf{v}_0^{(i)T})^T\right)$ which is uncorrelated with $\mathbf{Y}_{h-1,t-1}$ (i.e. $y_{it}=\mathbf{v}_0^{(i)T}\mathbf{z}_t$ follows an SCM(v_i,v_i) – see §5.3.4 and §5.3.5). $\mathbf{V}_s^{(i)}$ corresponds to a row of Ξ_s. The zero entries in $\mathbf{v}_j^{(i)}$ can be identified from the row dependencies in H_h. For the example above, we have (in order of increasing KI)

$$\mathbf{v}_0^{(2)}=(v_{01}^{(2)},1,0,0,0)^T,$$

$$\mathbf{v}_0^{(3)}=(v_{01}^{(3)},0,1,0,0)^T,\ \mathbf{v}_1^{(3)}=(v_{11}^{(3)},0,v_{13}^{(3)},v_{14}^{(3)},v_{15}^{(3)})^T,$$

$$\mathbf{v}_0^{(1)}=(1,0,0,0,0)^T,\ \mathbf{v}_1^{(1)}=(v_{11}^{(1)},0,0,v_{14}^{(1)},v_{15}^{(1)})^T,\ \mathbf{v}_2^{(1)}=(v_{21}^{(1)},0,v_{23}^{(1)},v_{24}^{(1)},\ v_{25}^{(1)})^T$$

and so on (the pattern of zero and non–zero entries is obtained from D in [5.51]. In terms of non–zero entries, row i of Φ_j ($[\Phi_j]_{i.}$), $j>0$ is given by row (v_i-j) of D and $[\Phi_0]_{i.}=(D(v_i,1),\ldots,D(v_i,i-1),1,0,\ldots,0)$, so that in Φ_0, the lower triangular part of row i is given by the dependencies in $D_{v_i.}$ **before** entry i). Ξ_s can now be assembled by placing the vectors $\mathbf{V}_s^{(i)T}$ on top of each other in some order. For instance, in the example above we take $s=\max(v_i)=4$ and

$\Xi_s =$

Φ_4	Φ_3	Φ_2	Φ_1	Φ_0
0 0 0 0 0	0 0 0 0 0	X 0 X X X	X 0 0 X X	1 0 0 0 0
0 0 0 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0	X 1 0 0 0
0 0 0 0 0	0 0 0 0 0	0 0 0 0 0	X 0 X X X	X 0 1 0 0
X 0 X X X	X 0 0 X X	0 0 0 X 0	0 0 0 X 0	0 0 0 1 0
0 0 0 0 0	0 0 0 0 0	X 0 X X X	X 0 0 X X	0 0 0 X 1

Since $\Xi_s \mathbf{F}_{s,t}$ is uncorrelated with $\mathbf{Y}_{h-1,t-1}$, $\Xi_s \mathbf{F}_{s,t}$ follows a VMA process of order less than s. Equivalently $\mathbf{z}_t$ follows a VARMA(s,s) model, although Φ_0 does not necessarily equal the identity matrix (as we see from the example). Within the VARMA(s,s) model, we can also make some simplifications in the MA structure, since for a component with KI ν_i, the associated linear combination $y_{i,t+\nu_i}$ (see above) is uncorrelated with $\mathbf{Y}_{h-1,t-1}$ and hence requires MA parameters in at most ν_i lags. The simplification is then that $[\Theta_j]_{i.} = \mathbf{0}$, $j > \nu_i$.

These considerations enable us to specify a sparse VARMA structure for our data with a possible saving in parameters. Notice though that we still have a contemporaneous coefficient matrix Φ_0 which must be multiplied out as suggested in [5.7] or [5.8] of §5.2.1 and we may lose the detailed sparsity **within** the rows by doing this (although zero rows in each Φ_j will be preserved when using the form [5.7]). We are free to arrange the rows of Ξ_s in any particular order, since these correspond to k simultaneous equations. Notice that the KIs (the number of non–zero (block) rows in Ξ_s) remain tied to a particular equation, so that we can reorder the columns of each Φ_j matrix (and so the order of the components in $\mathbf{z}_t$), but this does not alter any zero rows of the Φ_j matrices. Solo (1986) proves that the set of KIs $\{\nu_i\}$, are unique and that the "row degrees" of Ξ_s (the number of non–zero blocks in each row of Ξ_s) are invariant to multiplication of Ξ_s by a unimodular matrix (a matrix polynomial with constant, non–zero determinant – such as a reordering transformation).

5.3.2 Finding KIs

As demonstrated in §5.3.1, the KIs of a vector process can be determined by examining the row linear dependencies of a certain constructed covariance matrix H_h. Equation [5.48] shows that this matrix is derived from the constructed past and future variables, $\mathbf{Y}_{h-1,t-1}$ and $\mathbf{F}_{h-1,t}$, respectively. One way to explore the dependencies is with canonical correlation analysis, in which case a dependent row corresponds to a zero canonical correlation between the variables, and the canonical vector α, associated with $\mathbf{F}_{h-1,t}$ (see §5.2.8) will determine the nature of the linear dependency (e.g. which rows are included), specifically $\alpha_i = \mathbf{V}_s^{(i)}$ from above (when the i'th KI is found; which can be used to obtain initial estimates of the autoregressive parameters – see the example in §5.3.3 below).

T89a also details the modified statistic $C(s)$ of [5.36] (with "m" restricted to equal "j"; he also describes the *DIC* statistic of Akaike (1976), together with a suitable modification) for testing zero canonical correlations, which overcomes the problem due to the serial dependency inherent in the data.

Since we are searching for minimal KIs, we examine the dependencies from the top (of H_h) down and so construct subvectors $\mathbf{F}_t^*$ of $\mathbf{F}_{h-1,t}$ (by sequentially adding terms of $\mathbf{F}_{h-1,t}$ into the subvector) to use with $\mathbf{Y}_{h-1,t-1}$. When a zero row is found (e.g. row 2 of the example above), we leave out this (z_{2t}) and the corresponding future entries ($z_{2,t+j}$) from $\mathbf{F}_t^*$, so as to avoid unnecessary duplication. For the example above we first take $\mathbf{F}_t^*=(z_{1t})$, which is linearly independent of $\mathbf{Y}_{h-1,t-1}$, then we take $\mathbf{F}_t^{*T}=(z_{1t},z_{2t})$ at which point a zero canonical correlation will be found, so that $z_{2,t+j}$ is left out of the subvector. Subsequently we take $\mathbf{F}_t^{*T}=(z_{1t},z_{3t})$ (independent), followed by (z_{1t},z_{3t},z_{4t}) etc.

T89a attributes the KI approach to Akaike (e.g. 1976) and subsequently Cooper and Wood (1982). We will compare this approach with that of T&T in §5.5

5.3.3 Example of the KI approach

We will again use the Lydia Pinkham data ($\tilde{\mathbf{z}}_t$, normalised as in §5.2.10) for illustration and apply T89a's KI method to build a VARMA model for it.

Constructing the past vector $\mathbf{Y}_{h-1,t-1}$ with h=3 (since we suspected an overall order of 2 in the previous analysis of §5.2.10), the sequence of subvectors $\mathbf{F}_t^*$, and examining the canonical correlations between the past and future vectors in order to study the row dependencies of H_h, we obtain the following (from an implementation of the method in the statistical package S (S–plus (1988)).

Row	$\mathbf{F}_t^{*T}$	Corr	Stat	p–value	Dept? (KI)
1	(z_{1t})	0.70	33.36	<0.01	NO
2	(z_{1t},z_{2t})	0.35	6.66	0.25	YES (ν_1=0)
3	$(z_{1t},z_{1,t-1})$	0.58	20.63	<0.01	NO
5	$+z_{1,t-2}$	0.48	12.70	0.01	NO
7	$+z_{1,t-3}$	0.37	3.00	>0.25	YES (ν_2=3)

That is, the row dependencies in H_h have the form given in the following table and the KIs are 0 and 3.

"lag"	comp 1	comp 2
0	X	0
1	X	0
2	X	0
3	0	0

These dependencies allow us to specify a row sparse VARMA(3,3) model with parameters to be estimated in positions marked "X"

Φ_3		Φ_2		Φ_1		Φ_0		Θ_3		Θ_2		Θ_1	
X	0	X	0	X	0	1	0	X	X	X	X	X	X
0	0	0	0	0	0	X	1	0	0	0	0	0	0

The two canonical vectors α_i (such that $\alpha_i^T \mathbf{F}_{h-1,t}$ are each uncorrelated with $\beta_i^T \mathbf{Y}_{h-1,t-1}$, for the canonical vectors β_i) are

$$\alpha_1^T = \mathbf{V}_3^{(1)T} = (-0.06, 0, 0.11, 0, -0.07, 0, -0.12, 0),$$

$$\alpha_2^T = \mathbf{V}_3^{(2)T} = (0, 0, 0, 0, 0, 0, -0.11, 0.17),$$

which suggest initial estimates of the AR parameter matrices as

$$\Phi_0^{[S]} = \begin{bmatrix} 1 & 0 \\ -0.65 & 1 \end{bmatrix}, \Phi_1^{[S]} = \begin{bmatrix} 0.54 & 0 \\ 0 & 0 \end{bmatrix}, \Phi_2^{[S]} = \begin{bmatrix} -0.86 & 0 \\ 0 & 0 \end{bmatrix}, \Phi_3^{[S]} = \begin{bmatrix} 0.46 & 0 \\ 0 & 0 \end{bmatrix}$$

(see model [5.6]). We can thus build a VARMA(3,3) model for the transformed data $T\tilde{\mathbf{z}}_t$ in the form of [5.7], with initial estimates

$$T = \Phi_0^{[S]}, \Phi_1^{[T]} = \begin{bmatrix} -0.54 & 0 \\ 0 & 0 \end{bmatrix}, \Phi_2^{[T]} = \begin{bmatrix} 0.86 & 0 \\ 0 & 0 \end{bmatrix}, \Phi_3^{[T]} = \begin{bmatrix} -0.46 & 0 \\ 0 & 0 \end{bmatrix}.$$

This is not the same as model [5.39], built using T&T's method in §5.2.10, but, until we actually estimate both models, we cannot really make comparisons (see §6.1).

5.3.4 New definition of SCM

T89a gives a new definition of SCMs, but does not show that it is equivalent to that of T&T. A SC $y_t = \mathbf{v}_0^T \mathbf{z}_t$ is said to follow an SCM(p,q) if $u_t = y_t + \sum_{i=1}^{p} \mathbf{v}_i^T \mathbf{z}_{t-i}$ is

(i) uncorrelated with ψ_{t-j} for **any** $j>q$ and

(ii) correlated with ψ_{t-q}.

The definition in §5.2.2 used the conditions that u_t should be

(iii) uncorrelated with $\mathbf{a}_{t-j}$ for **all** $j>q$ and

(iv) correlated with $\mathbf{a}_{t-q}$.

To see the equivalence of these two definitions,

(i) certainly holds $\Longleftrightarrow$ u_t is uncorrelated with each member of $\{\mathbf{a}_{t-q-1}, \mathbf{a}_{t-q-2}, \ldots\}$ i.e. $\Longleftrightarrow$ (iii), and given that (i) holds then

(ii) holds $\Longleftrightarrow$ u_t is correlated with $\mathbf{a}_{t-q}$ $\Longleftrightarrow$ (iv).

Hence the two definitions are identical. □

Eventual Forecasting Components

T89a also defines the SC y_t to be an **Eventual Forecasting Component** (EFC) of $\mathbf{z}_t$ with index v if v is the smallest non–negative integer such that [5.47] holds (where now $\mathbf{v}_0$ is not necessarily in the form of the $\mathbf{v}_0$ of [5.47]). T89a then shows that the KIs of a vector process are precisely the indices of the k EFCs.

5.3.5 Comparison of SCMs and KIs

T89a shows that an SCM(p,q) gives us an EFC(max(p,q)) and also that an EFC(v) will give rise to an SCM(p,q) such that $v \geq \max(p,q)$, but the proofs given are too brief. To prove the first, let $y_t = \mathbf{v}_0^T \mathbf{z}_t$ follow an SCM(p,q), say, i.e.

$$y_t + \sum_{i=1}^{p} \mathbf{v}_i^T \mathbf{z}_{t-i} = \sum_{i=0}^{q} \mathbf{u}_i^T \mathbf{a}_{t-i}, \text{ (where } \mathbf{u}_0 = \mathbf{v}_0),$$

then

$$\begin{aligned} y_{t+v|t-1} &= E[y_{t+v} \,|\, \Psi_{t-1}] \\ &= \sum_{i=0}^{q} \mathbf{u}_i^T E[\mathbf{a}_{t+v-i} \,|\, \Psi_{t-1}] - \sum_{i=1}^{p} \mathbf{v}_i^T E[\mathbf{z}_{t+v-i} \,|\, \Psi_{t-1}]. \end{aligned} \qquad [5.54]$$

Now for $v \geq s = \max(p,q) \geq q \geq i$, $E[\mathbf{a}_{t+v-i} \,|\, \Psi_{t-1}] = E[\mathbf{a}_{t+v-i}] = \mathbf{0}$ (since $\mathbf{a}_{t+v-i}$ is certainly independent of Ψ_{t-1} for $i \leq v$), so that $y_{t+v|t-1}$ is a linear combination of $\mathbf{z}_{t+v-1|t-1}, \ldots, \mathbf{z}_{t|t-1}$ (in fact of $\mathbf{z}_{t+v-1|t-1}, \ldots, \mathbf{z}_{t+v-p|t-1}$). Also, if $v < s$ there are three situations

(i) $p = q = s$ $\Rightarrow$ in [5.54] that we at least have terms in $\mathbf{a}_{t-1|t-1}$ and $\mathbf{z}_{t-1|t-1}$

(ii) $p < q = s$ $\Rightarrow$ in [5.54] that we at least have terms in $\mathbf{a}_{t-1|t-1}$

(iii) $q < p = s$ $\Rightarrow$ in [5.54] that we at least have terms in $\mathbf{z}_{t-1|t-1}$

and hence in all three cases, $y_{t+v|t-1}$ is **not** a linear combination of just $\mathbf{z}_{t+v-1|t-1}, \ldots, \mathbf{z}_{t|t-1}$, so that s is the minimum non–negative integer such that the condition in [5.48] holds and hence y_t is an EFC(s).

To prove the converse, let $y_t = \mathbf{v}_0^T \mathbf{z}_t$ be an EFC(v) which implies that (a) $y_{t+v|t-1}$ is a linear combination of $\mathbf{z}_{t+v-1|t-1}, \ldots, \mathbf{z}_{t|t-1}$ and (b) $y_{t+v-1|t-1}$ is not a linear combination of $\mathbf{z}_{t+v-2|t-1}, \ldots, \mathbf{z}_{t|t-1}$. If we let the linear combination be given by $y_{t+v|t-1} = -\sum_{i=1}^{v} \mathbf{v}_i^T \mathbf{z}_{t+v-i|t-1}$ (where we may have $\mathbf{v}_i = \mathbf{0}$, for some i) and let $u_t = y_t + \sum_{i=1}^{v} \mathbf{v}_i^T \mathbf{z}_{t-i}$ then we can expand $y_{t+v|t-1}$ to give us $E[u_{t+j} \,|\, \Psi_{t-1}] = 0$, $j \geq v$.

Now from the proof of Definition 1 of T&T in §5.2.2, we have that

$$u_t = \sum_{i=0}^{v} \mathbf{v}_i^T \mathbf{z}_{t-i} = \sum_{k=0}^{\infty} \mathbf{h}_k^T \mathbf{a}_{t-k} \text{ where } \mathbf{h}_k^T = \sum_{i=0}^{k} \mathbf{v}_i^T \Psi_{k-i}, \text{ (MA}(\infty) \text{ form, } \mathbf{v}_i = \mathbf{0},\ i > v),$$

$$\text{i.e. } E[\sum_{k=0}^{\infty}\mathbf{h}_k^T\mathbf{a}_{t+j-k}\,|\,\{\mathbf{a}_{t-1},\mathbf{a}_{t-2},\ldots\}]=0,\ j\geq v\ \Rightarrow \mathbf{h}_k=\mathbf{0},\ k>v$$

which truncates the sum to v terms. So we have that

$$y_t+\sum_{i=1}^{v}\mathbf{v}_i^T\mathbf{z}_{t-i}=\sum_{k=0}^{v}\mathbf{h}_k^T\mathbf{a}_{t-k}, \qquad [5.55]$$

i.e. y_t follows at most an SCM(v,v).

□

T89a only shows that $v\geq\max(p,q)$, but we show equality since from [5.55] we have that

$$E[y_{t+v-1}\,|\,\psi_{t-1}]=E[\sum_{k=0}^{v}\mathbf{h}_k^T\mathbf{a}_{t+v-1-k}-\sum_{i=1}^{v}\mathbf{v}_i^T\mathbf{z}_{t+v-1-k}\,|\,\psi_{t-1}]$$

$$=y_{t+v-1\,|\,t-1}=\mathbf{h}_v^T\mathbf{a}_{t-1\,|\,t-1}-\mathbf{v}_v^T\mathbf{z}_{t-1\,|\,t-1}-\sum_{i=1}^{v-1}\mathbf{v}_i^T\mathbf{z}_{t+v-1-i\,|\,t-1} \qquad [5.56]$$

and as y_t is an EFC(v), property (b) above implies that in [5.56] either $\mathbf{h}_v\neq\mathbf{0}$ or $\mathbf{v}_v\neq\mathbf{0}$ and hence y_t follows an SCM(p,q) such that $\max(p,q)=v$.

□

5.3.6 Minimal orders

T89a claims that if the SCM orders are minimal, then the EFC index will be minimal (so that this is the KI of the component). However, it is not clear that this is the case, since the SCM method cannot choose between, for example, an SCM(2,0) and an SCM(1,1) (since the sum of the orders is 2 in each case), whereas they have KIs 2 and 1 respectively suggesting that the (1,1) model would be more desirable for the KI method. (KIs choose p,q to minimise $\max(p,q)$, whereas SCMs choose them to minimise $p+q$.) We will consider this again in §5.5 and look at the problem of deciding between orders in §6.1 and §6.2.

The KI method forces the AR and MA orders to be equal, which may not always be the case and could lead to overparameterisation.

5.4 Tsay's (1989b) refined KI method

Tsay (1989b) (=T89b) also describes the KI method of T89a (attributed to Cooper and Wood (1982)), together with the modified testing statistic (and some simulations demonstrating its superiority to the original form). T89b then presents a refinement of the KI method which allows the AR and MA orders of the SCMs to be different, giving it greater flexibility and we will discuss this in the following sections. Comparisons with the methods of T&T and T89a will be made in §5.5.

5.4.1 Refinement of Past and Future vectors

§5.3.2 described the construction of the past and future vectors, $\mathbf{Y}_{h-1,t-1}$ and $\mathbf{F}_t^*$ (a subvector of $\mathbf{F}_{h-1,t}$), which are used to find the KIs of a system. If we find a zero canonical correlation when we include the term $z_{i,t+v}$, say in $\mathbf{F}_t^*$ (but **not** when we have only included terms up to $z_{i,t+v-1}$), this suggests that z_{it} follows an SCM(p,q) such that v=max(p,q) (§5.3.5). However, the KI method only suggests an SCM(v,v)

$$\sum_{i=0}^{v}\mathbf{v}_i^T\mathbf{z}_{t-i}=\sum_{i=0}^{v}\mathbf{u}_i^T\mathbf{a}_{t-i},\ \text{(where } \mathbf{u}_0=\mathbf{v}_0),$$

or equivalently

$$\xi_v^T\mathbf{F}_{v,t}=\upsilon_v^T\mathbf{B}_{v,t} \qquad [5.57]$$

(see [5.52], $\xi_v^T=(\mathbf{v}_v^T,\ldots,\mathbf{v}_0^T)$, $\upsilon_v^T=(\mathbf{u}_v^T,\ldots,\mathbf{u}_0^T)$). With the result of §5.3.5, an obvious question to ask is whether one of p and q is strictly less than v and we now show how the past and future vectors can be modified to test for this.

To test whether $p<v$

If $\mathbf{v}_i=\mathbf{0}$ for some $p<i\leq v$, then $\xi_p^T\mathbf{F}_{p,t+v-p}$ would be uncorrelated with $\mathbf{Y}_{h-1,t-1}$ and we could alter the future vector $\mathbf{F}_{t+j}^*$ accordingly to test for this (i.e. by removing the earliest elements, $\mathbf{z}_t$, then $\mathbf{z}_{t+1}$ etc.). The statistic for testing zero canonical correlations will have to be modified to take the j into account – it becomes $C(s)$ of [5.36]. The maximum j for which $\mathbf{F}_{t+j}^*$ is still uncorrelated with $\mathbf{Y}_{h-1,t-1}$ gives us the order $v-p$. If $j>0$ then this fixes $p=v-j$ and $q=v$, since max(p,q)=v.

To test whether $q<v$

If $\mathbf{u}_i=\mathbf{0}$ for some $q<i\leq v$, then

$$\xi_v^T\mathbf{F}_{v,t}=\upsilon_q^T\mathbf{B}_{q,t+v-q} \qquad [5.58]$$

and $\xi_v^T\mathbf{F}_{v,t}$ would be uncorrelated with $\mathbf{Y}_{h-1,t+v-q-1}$. We can thus alter the past vector $\mathbf{Y}_{h-1,t+j-1}$ and perform a canonical correlation analysis between this and our previous future subvector, $\mathbf{F}_t^*$ to check for this form of redundancy (although "past" and "future" no longer really apply, since the vectors can contain common elements). The maximum j for which we still find a zero canonical correlation gives us $v-q$.

T89b suggests checking if $p<v$ first, so that the future vector $\mathbf{F}_t^*$ is fixed, although it is not clear that this is essential.

5.4.2 Example of the refined procedure

We will demonstrate the use of this refined procedure on the Lydia Pinkham data of §5.2.10 and §5.3.3. Since the procedure is a refinement of the KI method of §5.3, we continue the analysis of §5.3.3 having chosen KIs of 0 and 3. The component with KI of 0 cannot be simplified further, but perhaps that with KI 3 can be. In §5.3.3 we found a zero canonical correlation between $\mathbf{Y}_{h-1,t-1}$ and $\mathbf{F}_t^{*T}=(z_{1t},z_{1,t+1},z_{1,t+2},z_{1,t+3})$ (i.e. $\mathbf{F}_t^*$ includes "lags" of 0,1,2 and 3). The table below summarises the results of deleting elements from the future subvector.

j	$\mathbf{F}_{t+j}^*$ lags	Corr	Stat	p–value	Dept?
0	0,1,2,3	0.37	3.00	>0.25	YES
1	1,2,3	0.43	7.15	0.13	YES
2	2,3	0.51	6.91	0.23	YES
3	3	0.52	6.16	>0.25	YES

These results suggest that for this component, p=0 since we can decrease the future subvector and retain the linear dependency as suggested in §5.4.1 – i.e. we have an SCM(0,0) and an SCM(0,3), which considerably reduces the number of parameters required when compared with the results of the original KI analysis of §5.3.3. The transformation of the data necessary to exploit this parameterisation is the T of §5.3.3.

5.5 Comparison of approaches

We will now compare the various approaches to VARMA model specification discussed in the preceding sections, by examining the covariance structure upon which the methods rely, in order to see how the procedures work and how they are related. Throughout we will take $\Gamma_j = E[\mathbf{z}_t\mathbf{z}_{t+j}^T]$ (see §2.2.1), which is the most usual definition of the covariance function of a vector time series (**except** in T89a where he defines it to be Γ_{-j} for convenience of notation. The package S also uses Γ_{-j}).

5.5.1 KI method

Equation [5.52] of §5.3.1 rewrites the usual VARMA model in terms of constructed "future" vectors $\mathbf{F}_{s,t}$ and $\mathbf{B}_{s,t}$. Equation [5.53] then gives the derived covariance matrix H_h of [5.48] used in the KI approach. To determine the KIs we examine the row linear dependencies of H_h, excluding rows which are known to be dependent (as outlined in §5.3.2). In the first column of Table 5.4 below we display the submatrices of H_{h+1}^T which we would need to check for **column** linear dependencies, to search for KIs of increasing orders (we do this by looking for

right vectors – right eigenvectors corresponding to zero eigenvalues of the matrices). They are displayed as transposes of the usual form for more direct comparison with those matrices used by the SCM method (§5.5.2) and dependent columns due to components with KIs of lower orders have not been deleted. Notice also the different use of the parameter h in the methods which leads us to consider "h+1" for the KI based approaches.

Table 5.4 Covariance matrices used in model specification procedures.

KI(v)	SCM(m,j) (v=max(m,j))
v=0 $\begin{bmatrix}\Gamma_1\\ \Gamma_2\\ \vdots\\ \Gamma_{h+1}\end{bmatrix}$	(m,j)= (0,0) $\begin{bmatrix}\Gamma_1\\ \Gamma_2\\ \vdots\\ \Gamma_{h+1}\end{bmatrix}$
v=1 $\begin{bmatrix}\Gamma_1 & \Gamma_2\\ \Gamma_2 & \Gamma_3\\ \vdots & \vdots\\ \Gamma_{h+1} & \Gamma_{h+2}\end{bmatrix}$	(1,1) $\begin{bmatrix}\Gamma_2 & \Gamma_1\\ \Gamma_3 & \Gamma_2\\ \vdots & \vdots\\ \Gamma_{h+2} & \Gamma_{h+1}\end{bmatrix}$ (1,0) $\begin{bmatrix}\Gamma_1 & \Gamma_0\\ \Gamma_2 & \Gamma_1\\ \vdots & \vdots\\ \Gamma_{h+1} & \Gamma_h\end{bmatrix}$ (0,1) $\begin{bmatrix}\Gamma_2\\ \Gamma_3\\ \vdots\\ \Gamma_{h+2}\end{bmatrix}$
v=2 $\begin{bmatrix}\Gamma_1 & \Gamma_2 & \Gamma_3\\ \Gamma_2 & \Gamma_3 & \Gamma_4\\ \vdots & \vdots & \vdots\\ \Gamma_{h+1} & \Gamma_{h+2} & \Gamma_{h+3}\end{bmatrix}$	(2,2) $\begin{bmatrix}\Gamma_3 & \Gamma_2 & \Gamma_1\\ \Gamma_4 & \Gamma_3 & \Gamma_2\\ \vdots & \vdots & \vdots\\ \Gamma_{h+3} & \Gamma_{h+2} & \Gamma_{h+1}\end{bmatrix}$ (2,1) $\begin{bmatrix}\Gamma_2 & \Gamma_1 & \Gamma_0\\ \Gamma_3 & \Gamma_2 & \Gamma_1\\ \vdots & \vdots & \vdots\\ \Gamma_{h+2} & \Gamma_{h+1} & \Gamma_h\end{bmatrix}$ (2,0) $\begin{bmatrix}\Gamma_1 & \Gamma_0 & \Gamma_1^T\\ \Gamma_2 & \Gamma_1 & \Gamma_0\\ \vdots & \vdots & \vdots\\ \Gamma_{h+1} & \Gamma_h & \Gamma_{h-1}\end{bmatrix}$ (1,2) $\begin{bmatrix}\Gamma_3 & \Gamma_2\\ \Gamma_4 & \Gamma_3\\ \vdots & \vdots\\ \Gamma_{h+3} & \Gamma_{h+2}\end{bmatrix}$ (0,2) $\begin{bmatrix}\Gamma_3\\ \Gamma_4\\ \vdots\\ \Gamma_{h+3}\end{bmatrix}$

5.5.2 SCM method

Analogously to [5.52] we can write the VARMA(p,q) model of [2.2] in the form

$$\Phi\mathbf{Y}_{p,t}=\Theta\mathbf{A}_{q,t}, \qquad [5.59]$$

where

$$\Phi=\left[I:-\Phi_1:\cdots:-\Phi_p\right],\ \mathbf{Y}_{m,t}=(\mathbf{z}_t^T,\mathbf{z}_{t-1}^T,\ldots,\mathbf{z}_{t-m}^T)^T.$$

$$\Theta=\left[I:-\Theta_1:\cdots:-\Theta_q\right],\ \mathbf{A}_{m,t}=(\mathbf{a}_t^T,\mathbf{a}_{t-1}^T,\ldots,\mathbf{a}_{t-m}^T)^T.$$

If we postmultiply [5.59] by $\mathbf{Y}_{h,t-j-1}^T$ and take expectations we obtain

$$\Phi\begin{bmatrix}\Gamma_{j+1}^T & \cdots & \Gamma_{j+1+h}^T\\ \vdots & \ddots & \vdots\\ \Gamma_{j+1-p}^T & \cdots & \Gamma_{j+1+h-p}^T\end{bmatrix}=E[\Theta\mathbf{A}_{q,t}\mathbf{Y}_{h,t-j-1}^T]=0 \text{ if } j\geq q \qquad [5.60]$$

(since the newest element in $\mathbf{Y}_{h,t-j-1}$ is $\mathbf{z}_{t-q-1}$ and the oldest in $\mathbf{A}_{q,t}$ is $\mathbf{a}_{t-q}$). Taking the transpose of [5.60] gives

$$\begin{bmatrix} \Gamma_{j+1} & \cdots & \Gamma_{j+1-p} \\ \vdots & \ddots & \vdots \\ \Gamma_{j+1+h} & \cdots & \Gamma_{j+1+h-p} \end{bmatrix} \Phi^T=0 \Rightarrow \begin{aligned} \Gamma_{j+1} &= \Gamma_j\Phi_1^T+\ldots+\Gamma_{j+1-p}\Phi_p^T \\ &\vdots \\ \Gamma_{j+1+h} &= \Gamma_{j+h}\Phi_1^T+\ldots+\Gamma_{j+1+h-p}\Phi_p^T \end{aligned} \quad (j\geq q) \quad [5.61]$$

(note that the constructed covariance matrix in [5.61] is precisely $\Gamma(p,h,j)$ of T&T). These are the usual recursions for the covariance matrices of a VARMA(p,q) model when $j\geq q$ and the so MA terms do not have any effect (see e.g. Jenkins and Alavi (1981)). From these we could (theoretically at least, provided that $h\geq p$ to give us enough simultaneous equations) calculate the orders p and q and estimates of the AR parameter matrices Φ_i, $i=1,..,p$. T&T's method searches for right vectors $\mathbf{v}$, of $\Gamma(m,h,j)$ for increasing m and j in some order, which correspond to rows of Φ. The second column of Table 5.4 shows the matrices $\Gamma(m,h,j)$ used to find SCMs of various orders.

5.5.3 T89b's refined KI method

T89b's refinement to the KI method involves calculating the KIs of a system and then refining the past and future vectors to search for further simplification in the AR or MA orders of the components. To test if the AR order $p<v$ (the KI) of a component, we first remove the oldest elements of the future vector, which removes the block column $[\Gamma_1^T:\cdots:\Gamma_{h+1}^T]^T$ from H_{h+1}^T to give $\Gamma(v-1,h,v)$ of T&T (with the block columns reversed). e.g. see Table 5.4 $v=2$, $(m,j)=(2,2)$ then $(m,j)=(1,2)$. If this matrix still contains a dependent column, the procedure is repeated (e.g. $(m,j)=(0,2)$), otherwise the past vector is lagged to give $\Gamma(v,h,v-1)$ (with the block columns reversed. e.g $(m,j)=(2,2)$, $(2,1)$), which is then tested for column linear dependencies.

5.5.4 Relationships between the methods

It can be seen that the KI method uses the same covariance matrices as T&T (with the block columns reversed) with $m=j=v$ and T89b's refinement then uses the same matrices as T&T, first trying $m<v$ and then $j<v$. However T89a and T89b both add one column at a time to the covariance matrix, whilst T&T adds a block column at a time. Clearly the KI method is restrictive whilst both T&T and T89b allow all possible model orders to be covered. An important difference is that T89b deletes dependent rows/columns from future consideration which avoids the problem of redundant right vectors which T&T still has to deal with (see §5.2.9).

5.5.5 Search path

We can calculate the number of "searches" (or canonical correlation analyses) required by each of the methods to find an SCM(m,j) by looking at the search paths. The numbers required by T&T ($=N_{T\&T}(m,j)$), using their suggested "parsimonious" search path is given in [5.38] of §5.2.9. If we had found any SCMs of orders lower than those we were currently considering, then we would have to perform an additional analysis at each higher order to determine which right vectors were due to the lower order SCMs. Table 5.5a below, modifies the number of searches to take this into account.

Table 5.5 Number of searches to find an SCM(m,j).

(a) $N_{T\&T}$

$m\backslash j$	0	1	2	3	4	5
0	1	3	7	13	21	31
1	5	9	15	23	33	45
2	11	17	25	35	47	61
3	19	27	37	49	63	79
4	29	39	51	65	81	99
5	41	53	67	83	101	121

(b) N_{T89b}

$m\backslash j$	0	1	2	3	4	5
0	1	3	5	7	9	11
1	4	4	5	7	9	11
2	6	6	5	6	8	10
3	8	8	7	6	7	9
4	10	10	9	8	7	8
5	12	12	11	10	9	8

The number grows large even for modest orders. The number of searches required by T89b's refined method to find an SCM(m,j) ($=N_{T89b}(m,j)$) are given in Table 5.5b. To see how the entries arise, notice that to confirm SCMs of the orders shown in the following table, we must also check the other orders shown (we also display their dependencies, assuming that the SCM is of the order given).

SCM	Also check orders					
(1,2)	(0,0)	(1,1)	(2,2)	(1,2)	(0,2)	
Dept?	X	X	0	0	X	
(2,1)	(0,0)	(1,1)	(2,2)	(1,2)	(2,1)	(2,0)
Dept?	X	X	0	X	0	X
(2,2)	(0,0)	(1,1)	(2,2)	(1,2)	(2,1)	
Dept?	X	X	0	X	X	

The original KI method of T89a cannot find an SCM(m,j) for $m \neq j$, but it takes the same number of searches as T89b to find an SCM(m,m). We see that T&T requires more searches than T89b to find models partly because it has to deal with exchangeability and also because the path over the root table is not as efficient. The saving when using T89b is quite substantial for larger order SCMs.

For k SCs, T&T considers all k simultaneously at each order (m,j) (with one canonical correlation analysis) while T89b considers each SC (with an unknown order at that stage) separately. Thus when modelling a k dimensional process, the searches in Table 5.5a remain the same, while those in Table 5.5b will increase with k depending on the exact SCM structure (although now the canonical

correlation analyses are not comparable between the tables – those in 5.5a are perhaps k times as "large" as those in 5.5b). For processes which can be adequately modelled by low order SCMs ($m,j \leq 1$, say), T&T may find these quicker than T89b.

5.5.6 Recommendations

With the considerations above, it can be seen that T89a does not have the flexibility of the other two methods and is a special case of T89b, so that it is likely to be preferable to use T89b. It appears that T89b may be preferable to T&T for larger order models since it produces the same result (using the same covariance matrices), but may be able to arrive there with less canonical correlation analyses by using a more efficient search path. This is also helped by the saving made from not introducing redundant rows in the covariance matrices. We consider some further aspects of the methods of T&T and T89b in Chapter 6, in particular their ways of dealing with exchangeable models – see §6.5.

5.6 Other methods

In preceding sections we have looked in some detail at three related methods of VARMA model specification. In this section we will briefly discuss some other possible approaches to the problem. (Other models, such as those mentioned in §2.6 will require different methods to specify them and we will not consider any of these, since we have not discussed the models in sufficient detail.) Many papers deal with the univariate model specification problem (see for example, Gooijer et. al. (1985) and Piccolo and Tunnicliffe Wilson (1984) for surveys of the methods), although as mentioned in §5.1, multivariate generalisations must include ways of reducing the number of parameters and not just of identifying possible model orders. We now consider some approaches to the multivariate problem, grouped together into those with similar objectives.

5.6.1 Dimension reduction

Peña and Box (1987) describe a factor analytic method for reducing the dimension (number of series) of a multivariate time series problem. It searches for underlying factors (combinations of the original data) which are thought to drive the observed series, hoping that only a few of these are needed for an adequate representation. Principal component methods (see e.g Priestley (1981)) present a similar solution. The parameter reduction produced by such techniques could be achieved in a special case of T&T's SCM formulation (where we have several SCM(0,0)s) and application of T&T (or T89b) may serve to highlight this.

Box and Tiao (1977) decompose a (stationary, multivariate) series into components in increasing order of predictability which essentially orders them in increasing degrees of non–stationarity. Again, such a decomposition may sometimes be useful and the need for it may be indicated by application of some other procedure. For n.s. series, such a decomposition may find CI relationships – see §5.2.12.

Shapiro and Switzer (1989) decompose series into components with increasing autocorrelation (similarly to Box and Tiao (1977)) in search of common trends among series and to allow only the "noise" (weakly autocorrelated) components to be smoothed (which can give a smoother version of the data). These two applications appear to be useful with large numbers of noisy series.

A dimension reduction technique will be sensible when this is thought to be a plausible feature of the data. However, without such prior opinions a more flexible approach, such as T&T or T89b could be used to indicate whether this were likely to be the case (if SCM(0,0)s can be found) and then, perhaps a dimension reduction could be sought using one of these more specialised methods. It may prove worthwhile to investigate the application of T&T to datasets which are suitable for dimension reduction.

5.6.2 Parameter reduction

Koreisha and Pukkila (1987) describe a method for obtaining initial estimates of the parameters in a VARMA model, hopefully enabling us to delete many parameters from a representation before we apply a more accurate (and so, time–consuming) estimation procedure. However, parameter redundancy such as in §5.1.1 can still give rise to significant coefficients and the transformation to be applied will allow us to exploit this, which simply identifying insignificant parameters will not.

Bayesian vector autoregression (e.g. Litterman (1986)) imposes prior beliefs about the coefficients in a VAR model for the data, which encourage parsimonious parameterisations. A major application of these models is in forecasting where they have been shown to perform well due to the smaller number of parameters, however they cannot share the descriptive flexibility of VARMA models. The intended use of such a model must be a major consideration.

Young et. al. (1980) discuss the "Instrumental Variable" approach to model order identification which recursively produces estimates of model parameters (in their examples they use single–input single–output transfer function noise models, although the method could be extended to VARMA models; it is a natural

technique to use for recursive time series modelling (Young (1984)) of time–varying or possibly non–linear systems) and provides a measure of overparameterisation based upon the variance of the estimated model parameters. The method chooses models which both fit the data well (have low residual error variance) and are believed to be parsimonious since the parameter estimates are precise. i.e. adding more parameters will certainly decrease the residual error variance, but it may also increase the variance of the estimated parameters and if this increase is "large" compared with the decrease in residual variance, then the extra parameters are taken to be redundant. In the presence of redundancy such as that discussed in §5.1.1 it may also be necessary to consider the **covariance** between the parameter estimates

5.6.3 Parsimonious specification

Ahn and Reinsel (1988) consider VAR models with the coefficient matrices of reduced rank (as in the example model of §5.1.1), so that we can factorise out a matrix T (this can be estimated by maximum likelihood along with the other model parameters in their model representation). However, the redundancy is assumed to be **nested**, so that the rows of $\Phi_{i+1}^{[T]}$ (see [5.7]) are linearly dependent on those of $\Phi_i^{[T]}$. This is not necessarily true with T&T when we have MA terms. Ahn and Reinsel's method uses the same covariance matrices to search for the redundancy and so produces the same result as T&T constrained to search over orders with q=0.

Akaike (e.g. 1976) and subsequently Cooper and Wood (1982) apply canonical correlation analysis to help specify VARMA(p,p) models. This approach is essentially the same as that of T89a (§5.3) who also derives more consistent statistics to use.

5.6.4 Conclusion

The methods of T&T and T89b can be useful tools for specifying VARMA models. Application of such procedures may also highlight features of the data which can suggest further or different analyses and the parsimonious parameterisation should give more stable models from which to make deductions. Whether they will give more accurate forecasts remains to be seen. We will assess the models produced by T&T in Chapter 6 and compare them with those produced by the more usual techniques of Chapter 2.

Chapter 6. Case studies

In this chapter we will build VARMA models for some published datasets using both the "standard" model identification tools of Chapter 2 and the more sophisticated model specification procedures of Chapter 5, as well as investigating co–integration and the effects it has on the models, using the techniques of Chapter 4. We can check the fit of the models using the diagnostic checking outlined in §2.4 and compare the models using criteria of §3.3, in order to assess whether the model specification procedures are worth applying to the chosen datasets, although we will not attempt to interpret the actual models obtained. The analyses will also serve to illustrate many of the points discussed in the preceding chapters.

Figures and Tables for this chapter will appear at the end of each section.

6.1 Annual Lydia Pinkham data

This dataset is described in Appendix A.1 where a brief background is given and previous analyses are discussed. In the following section we will apply the methods of Chapters 4 and 5 to see if their use is worthwhile on this dataset.

Figure A.1a plots the annual data from 1907 to 1960. Sales are larger than advertising and may even lead them a little as some studies have suggested. The interventions mentioned in §A.1 appear as change points in both series and may need correcting for. In their analysis, Heyse and Wei (1985b) present Tiao and Box's (1981) advice about dealing with non–stationarity in VARMA models – i.e. allowing the roots of the polynomial $|\Phi(B)|$ (the autoregressive operator in the VARMA model) to lie on the unit circle (§3.1.2). This does not always correspond to differencing each series and avoids any difficulties due to overdifferencing. Heyse and Wei suggest that both of the annual series require first differencing and claim that in their analysis they did not have any problems with unit roots in the moving average polynomial and hence did difference both series. These considerations are of course related to co–integration and so we will apply the techniques of §4.4 to see if the series are not CI.

6.1.1 CI analysis

Let S_t= sales at time t, A_t= advertising at time t. Figures 6.1a, b, d and e show the acf and pacf of S_t and A_t respectively (Figures 6.1c and f show the acf of ∇S_t and ∇A_t). These suggest that the series may be n.s. (as do the plots in Figure A.1a). Fitting univariate AR(1) models to each series gives $\hat{\varphi}_1$ parameters of 0.91 and 0.81 respectively, suggesting that differencing may be necessary. To check whether the series are CI, we regress S_t on A_t, the residuals from which are shown

in Figure 6.1g. This plot and their acf in Figure 6.1h suggests that the residuals are still n.s. The usual tests of §4.4 can be applied, but we must now use the critical values of Engle and Yoo (1987) (see §4.7.2) since we have fewer observations than were used in the original simulations of E&G. For the regression residuals, the statistic $\hat{DF}$ (which is thought to be the more robust test) =3.092 which is smaller than the 10% point of 3.28 for 50 observations of Engle and Yoo (1987). We conclude that the series are each n.s. and not CI, so that it should not give any problems due to overdifferencing to analyse $s_t=\nabla S_t$ and $a_t=\nabla A_t$, as Heyse and Wei found.

6.1.2 Model identification

Heyse and Wei chose to fit a VAR(2) model to the differenced series $\mathbf{z}_t=(a_t,s_t)^T$, based upon the cut–offs in the partial **autoregression** matrices (§2.2.2). We display these matrices in Table 6.1b, together with the cross correlation matrices in Table 6.1a (and the collected significances for each entry at all lags as suggested by Tiao and Box (1981)) and the partial lag correlation matrices of Heyse and Wei (1985a) in Table 6.1c. The cut–offs in these (measured by the statistics M and X of §2.2.2 (see equations [2.18] and [2.23]) – the p-values in Tables 6.1b and c refer to the upper percentage point of the relevant χ^2 distribution) suggest a VAR(2) model. We also display the significances of each entry of the ccf when compared with the accurate variance estimate of [2.14] in §2.2.1 and the conclusions drawn are very similar. Experience with other cases also suggests that $1/N$ is an adequate estimate of the variance to use in practice. We have re–estimated Heyse and Wei's (1985b) final VAR(2) model by exact maximum likelihood using NAG routine g13dcf (Numerical Algorithms Group (1988)) giving the following results.

$$\hat{\Phi}_1^{[1]}=\begin{bmatrix}-0.23 & 0.56\\ 0 & 0.48\end{bmatrix},\ \hat{\Phi}_2^{[1]}=\begin{bmatrix}-0.50 & 0\\ -0.24 & 0\end{bmatrix},\ \hat{\Sigma}_a^{[1]}=\begin{bmatrix}0.62 & 0.32\\ 0.32 & 0.74\end{bmatrix}. \qquad [6.1]$$

(Heyse and Wei estimate this model on only the first 39 observations, keeping the remaining 14 for forecasting – see §6.1.4). Model [6.1] was estimated with the normalised (zero mean, unit variance) data $\mathbf{z}_t$, (see §3.5.2(a) and §5.2.10) since the original scale is too large for the estimation routine to handle. In their analysis, Heyse and Wei fitted a sequence of 3 models, deleting insignificant parameters at each stage. In model [6.1], all estimated parameters are greater than twice their standard error (in modulus). "0" represents a value which has been held at zero during the estimation. Fitting information is summarised in Table 6.2, together with that from the other models below for comparison.

6.1.3 Diagnostic checking

To check the fit of model [6.1], we consider the residuals $\hat{\mathbf{a}}_t$, from it. The two component series are plotted in Figures 6.2a and b, together with $\pm 2\sqrt{[\hat{\Sigma}_a]_{ii}}$ limits to highlight outliers. The residuals are also plotted in the "uncorrelated" form of §2.4 in Figures 6.2c and d with $\pm 2\sqrt{\hat{\lambda}_i}$ limits (see §2.4). There do not appear to be any particularly large outliers in the residuals (the advertising residuals from 1918 and 1925 can perhaps be explained by the lags from sales seen in these years – see Figure A.1a) and they are uncorrelated – all of the correlation matrices (equivalent to those in Tables 6.1) have insignificant entries, suggesting that model [6.1] fits reasonably well.

6.1.4 Forecasting

Heyse and Wei estimate model [6.1] on only the first 39 observations and use this to forecast the last 14 (1947–1960). With only this small number of annual observations available, 14 years are probably too many to attempt to forecast, so we will hold back the last 3 years' data and re–estimate model [6.1] on the first 50 observations (the differenced data consists of 53 observations). We can then use this for a small evaluation of model [6.1]'s out–of–sample forecasting performance. All forecasting is done with the differenced, normalised data $\tilde{\mathbf{z}}_t$ and the results are summarised in Table 6.3 below. Also shown are the results from the other models for comparison.

6.1.5 T&T's method

Application of the method of T&T in §5.2.10 suggested that a row–sparse VAR(2) model might be suitable for some transformation, T of the data (see equation [5.39]). Estimating this model by exact maximum likelihood (using the values in equation [5.39] as initial estimates), we obtain (model [6.2(T)])

$$T=\begin{bmatrix}0.79 & -0.62\\ 0.62 & 0.79\end{bmatrix},\ \hat{\Phi}_1^{[2T]}=\begin{bmatrix}0 & 0\\ -0.63 & 0.46\end{bmatrix},\ \hat{\Phi}_2^{[2T]}=\begin{bmatrix}0 & 0\\ -0.43 & -0.29\end{bmatrix},\ \hat{\Sigma}_a^{[2T]}=\begin{bmatrix}0.42 & \\ 0.02 & 1.00\end{bmatrix}$$

which can be transformed back into a VAR(2) model for the original (differenced, scaled) data, $\tilde{\mathbf{z}}_t$ to give

$$\hat{\Phi}_1^{[2O]}=\begin{bmatrix}-0.13 & 0.46\\ -0.16 & 0.59\end{bmatrix},\ \hat{\Phi}_2^{[2O]}=\begin{bmatrix}-0.32 & 0.03\\ -0.41 & 0.03\end{bmatrix},\ \hat{\Sigma}_a^{[2O]}=\begin{bmatrix}0.66 & \\ 0.29 & 0.76\end{bmatrix}, \qquad [6.2(O)]$$

which is similar to model [6.1] (notice that we have some apparently small coefficients in model [6.2(O)], but these are linear combinations of significant, estimated parameters and cannot themselves be deleted). The residuals from this model also suggest that it fits reasonably well and modelling and forecasting results are summarised in Tables 6.2 and 6.3.

In §5.2.10 we had a choice of order for the second SC – (0,2) or (2,0). We chose the order (2,0) in order to compare the VARMA model with that of Heyse and Wei (model [6.1]). We now repeat the analysis, taking instead an SCM(0,2), to give us an overall VMA(2) model for $T\mathbf{z}_t$ with

$$T=\begin{bmatrix}-0.57 & 0.82\\ 0.82 & 0.57\end{bmatrix},\ \Theta_1^{[T]}=\begin{bmatrix}0 & 0\\ X & X\end{bmatrix},\ \Theta_2^{[T]}=\begin{bmatrix}0 & 0\\ X & X\end{bmatrix}$$

("X" indicates a parameter to be estimated, "0" a parameter held at zero during estimation – we do not have initial estimates of the MA parameters, as discussed in §5.2.9). Estimating this in terms of the transformed data and transforming back to the original data we obtain a VMA(2) model for $\mathbf{z}_t$ with

$$\hat{\Theta}_1^{[3]}=\begin{bmatrix}0.42 & -0.60\\ 0.29 & -0.42\end{bmatrix},\ \hat{\Theta}_2^{[3]}=\begin{bmatrix}0.58 & -0.03\\ 0.40 & -0.02\end{bmatrix},\ \hat{\Sigma}_a^{[3]}=\begin{bmatrix}0.62 & \\ 0.29 & 0.76\end{bmatrix}. \qquad [6.3]$$

Modelling information for this is also summarised in Table 6.2 and forecasting results in Table 6.3.

6.1.6 T89b's method

In §5.4.2, we applied the method of T89b to this dataset, which suggested a VMA(3) model. We have estimated this using exact maximum likelihood to give the following results (transformed back in terms of the original data, $\mathbf{z}_t$)

$$\hat{\Theta}_1^{[4]}=\begin{bmatrix}-0.25 & 0\\ -0.16 & 0\end{bmatrix},\ \hat{\Theta}_2^{[4]}=\begin{bmatrix}0.25 & 0\\ 0.16 & 0\end{bmatrix},\ \hat{\Theta}_3^{[4]}=\begin{bmatrix}0.40 & 0\\ 0.25 & 0\end{bmatrix},\ \hat{\Sigma}_a^{[4]}=\begin{bmatrix}0.83 & \\ 0.43 & 0.86\end{bmatrix}. \qquad [6.4]$$

6.1.7 Comparison

AIC values for each of the models considered have been calculated using both definitions [3.9] and [3.11] of §3.3.1. In counting the number of parameters (Npar) to penalise the likelihood with, we have ignored the "parameters" in the transformation matrices (T) since these are not "independently adjustable" (see §3.5.1(c)) by the estimation routine (in the same way that the "0"s which they give rise to in the coefficient matrices are fixed and cannot be adjusted by the routine). The effort expended to find the transformations should be taken into account when comparing the models, but probably not simply in the form of extra parameters.

The results in Table 6.2 suggest that model [6.1] (Heyse and Wei, VAR(2)) fits the data best, although model [6.3] (T&T, VMA(2)) is very similar. These both fit better than model [6.2] (T&T, VAR(2)) by an amount equivalent to about 3 parameters. Model [6.4] (T89b, VMA(3)) fits worse than the others by about 8 parameters. These results suggest that the model specification procedures of T&T

and T89b can produce models which are worse than we might otherwise choose.

In these cases it is seen that both forms of AIC give similar results since they differ by a constant (although we usually only use the form of equation [3.11] for VAR models). The forecast values, errors and their standard errors in Table 6.3 are very similar for all of the models with little to choose between them on such a small evaluation.

It appears that the model specification procedures of T&T and T89b cannot improve upon Heyse and Wei's VAR(2) model and they even produce some possible models which are worse than this. However, [6.3], for example has fewer parameters and was quicker to estimate than [6.1], particularly since [6.1] was the result of three estimations. It would be desirable to have an indication of which model is likely to be preferable beforehand, particularly when we are left with a choice of possible model orders, as in this case. This dataset is quite small and easily modelled using standard techniques, so we might expect the model specification procedures to do better in larger and more complex situations when the standard techniques cannot be usefully applied. T89b gives a worse result than T&T and the saving in effort due to the more efficient search path (discussed in §5.5.5) is not relevant with such small models. The nature of the search used by T&T, which highlights exchangeable models may make this procedure preferable, since in this case the two exchangeable models are quite different.

Table 6.1a Cross correlations.

Lag	ccf	$2/\sqrt{N}$ significance	Accurate significance
1	0.05 0.22 0.31 0.43	. . + +	. . + +
2	−0.40 −0.14 −0.15 0.10	− . . .	− . . .
3	−0.09 −0.06 0.07 0.07	. . . .	. . . .
4	0.43 0.16 0.28 0.11	+ . + .	+ . . .
5	0.04 −0.14 0.19 0.03	. . . .	. . . .

Significances for each entry over all lags

$2/\sqrt{N}$		Accurate	
.−.+.	.....	.−.+.	.....
+..+.	+....	+....	+....

Table 6.1b Partial autoregressions.

Lag	parm	Significance	*M*	p-value
1	−0.20 0.43 −0.04 0.45	. + . +	12.56	0.01
2	−0.49 −0.03 −0.30 0.08	− . − .	16.38	0
3	−0.28 0.33 −0.12 0.15	− + . .	−4.46	1
4	0.17 0.10 0.07 −0.05	. . . .	0.2	1
5	−0.23 0.26 −0.45 0.21	. . − .	−15	1

Table 6.1c Partial lag correlations.

Lag	plcm	$2/\sqrt{N}$ sig	*X*	p–value
1	0.05 0.22 0.31 0.43	. . + +	17.56	0
2	−0.53 −0.28 −0.32 −0.10	− − − .	25.06	0
3	−0.12 −0.04 0.21 0.09	. . . .	3.49	0.48
4	0.24 0.05 0.21 −0.01	. . . .	5.47	0.24
5	−0.11 −0.32 0.16 −0.02	. − . .	7.42	0.12

Table 6.2 Lydia Pinkham data model fitting information.

Model	LL	Npar	$AIC_{[3.9]}$	(min=0)	$AIC_{[3.11]}$	(min=0)
VAR(2) [6.1]	–123.4	5	256.7	0.0	–44.9	0.0
VAR(2) [6.2]	–127.7	4	263.5	6.8	–38.0	6.9
VMA(2) [6.3]	–125.7	3	257.4	0.7	–44.5	0.4
VMA(3) [6.4]	–133.8	3	273.7	17.0	–27.8	17.1

Key: LL= log of maximised likelihood function; Npar= number of independently adjustable parameters; see §3.3.1 for definitions of *AIC*; the column (min=0) gives the AIC values rescaled by subtracting the minimum value.

Figure 6.1

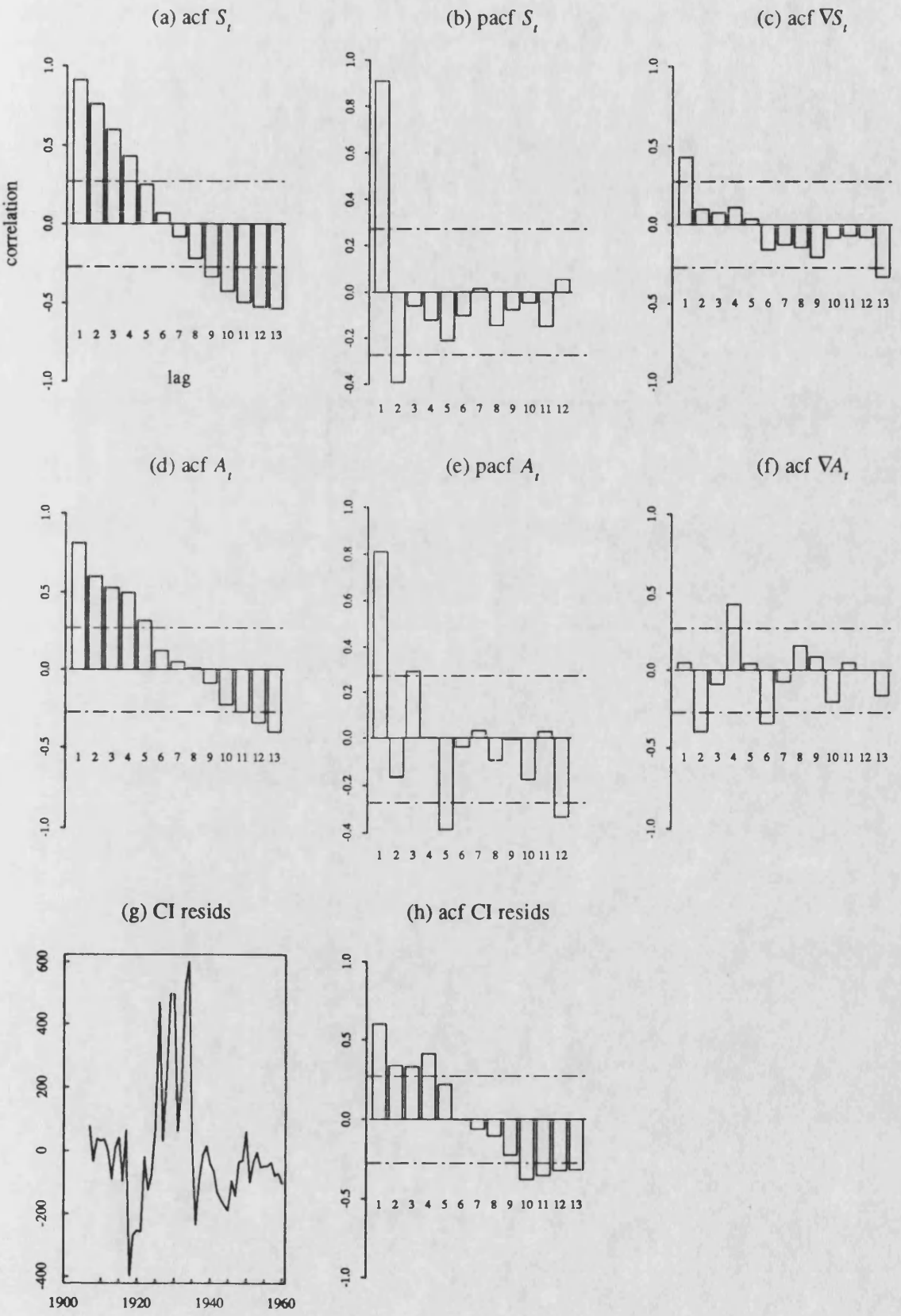

(a) acf S_t
(b) pacf S_t
(c) acf ∇S_t
correlation
lag
(d) acf A_t
(e) pacf A_t
(f) acf ∇A_t
(g) CI resids
(h) acf CI resids
1900
1920
1940
1960

Figure 6.2 Residuals from models

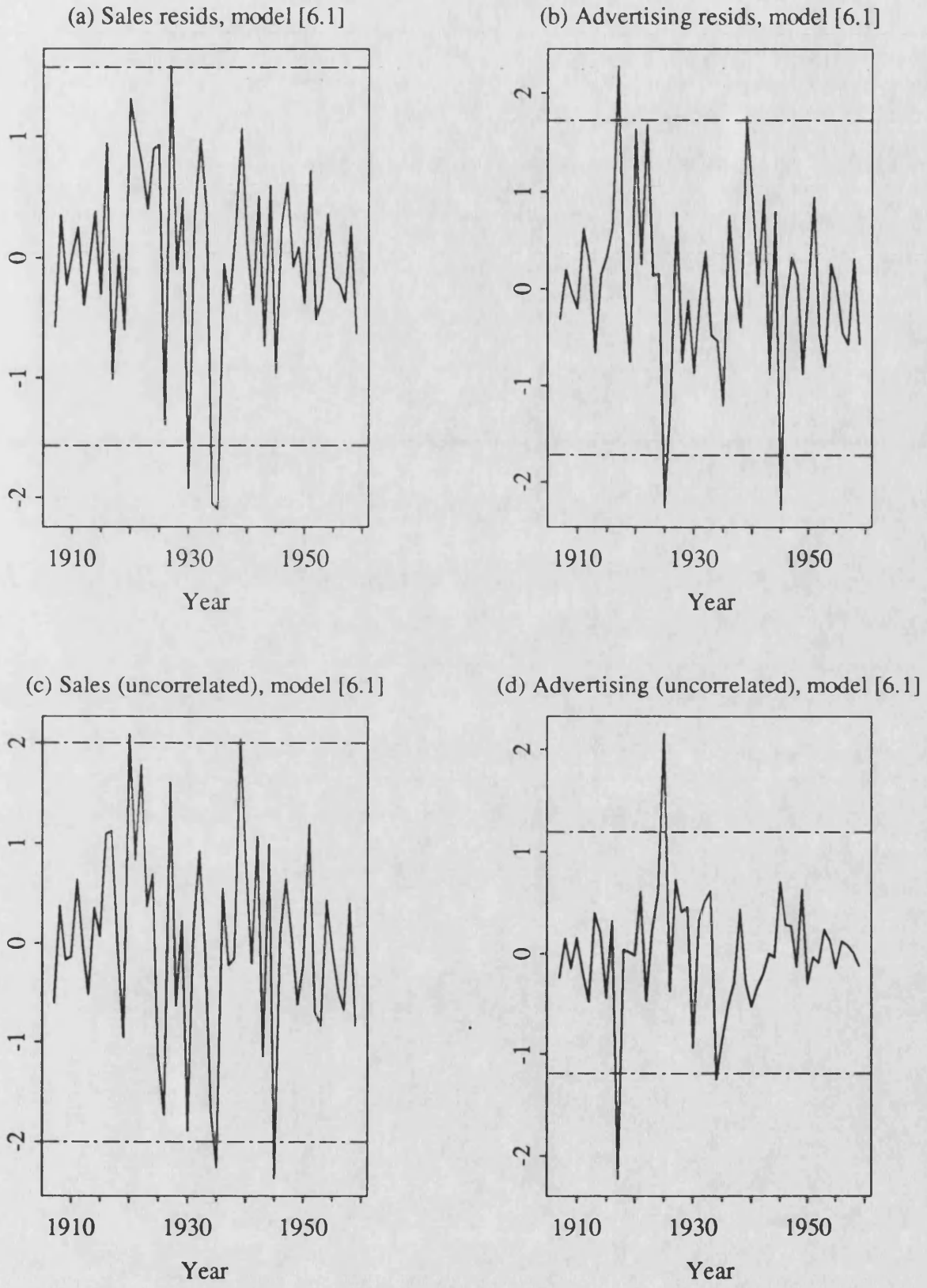


Table 6.3 Forecasts for Lydia Pinkham data.

Model	Period					
	1958		1959		1960	
	a_t	s_t	a_t	s_t	a_t	s_t
Actual	−0.59	0.78	0.03	−0.03	−0.36	−0.44
[6.1]	−0.22	−0.21	0.01	−0.06	0.08	0.03
Error	−0.37	0.99	0.02	0.03	−0.44	−0.47
(s.e.)	(0.80)	(0.88)	(0.91)	(0.97)	(0.97)	(1.00)
[6.2]	−0.18	−0.23	−0.05	−0.06	0.03	0.04
Error	−0.41	1.01	0.08	0.03	−0.39	−0.48
(s.e.)	(0.83)	(0.89)	(0.91)	(1.01)	(0.96)	(1.08)
[6.3]	−0.17	−0.11	0.03	0.02	0.00	0.00
Error	−0.42	0.89	0.00	−0.05	−0.36	−0.44
(s.e.)	(0.80)	(0.89)	(0.95)	(0.95)	(1.05)	(1.00)
[6.4]	−0.12	−0.08	0.19	0.12	0.11	0.07
Error	−0.47	0.86	−0.16	0.15	−0.47	−0.51
(s.e.)	(0.94)	(0.94)	(0.96)	(0.96)	(0.98)	(0.96)

6.2 Flour price data

This dataset is described in Appendix A.2 and in this section we will build some VARMA models for it using various techniques. The information criteria of §3.3.1 and some forecasting results can be used to compare the models and assess the usefulness of the model specification procedures. Throughout we will use the logged data as mentioned in Appendix A.2.

6.2.1 Co–integration

Each of the series are clearly n.s. (see the plots in Figure A.2, also the acfs decay slowly – e.g. Figure 6.3a) and require differencing. It may be expected that the proximity of the markets would keep a stable contemporaneous relationship between the series – i.e. they would be CI. Testing for this by regressing each of the series on the remainder (see §4.4.2) we find that each of the residual series so generated are themselves clearly n.s. (from the plots and acfs – see e.g. Figures 6.3b and c) and the usual tests confirm this as tabulated below (in this case we have the same number of observations as E&G used in their simulations (§4.4), so the critical values of Table 4.2 can be used).

Dependent variable	$\hat{DW}$ (10%=0.32)	$\hat{DF}$ (10%=2.84)	n.s.?
1: Buffalo	0.25	2.42	YES
2: Minneapolis	0.27	2.44	YES
3: Kansas	0.22	2.35	YES

These results suggest, perhaps rather surprisingly that the series are not CI and we can work with the differenced data to avoid complications with unit roots when fitting models. Some of the models we build with the aid of model specification procedures are in terms of the raw (undifferenced) data.

6.2.2 Model identification

The usual correlation matrices of Chapter 2 for the data $\nabla \mathbf{z}_t$ are tabulated in Tables 6.4. Notice that some of the entries of the parm at lag 1 are not correlations. The cut–offs in the "partial" functions are easier to spot in the plcm. These suggest that a VMA(1) or VAR(1) model may be appropriate. We can estimate these to give the following results.

Model [5]: $\nabla \mathbf{z}_t = (I - \Theta_1 B)\mathbf{a}_t$

$$\hat{\Theta}_1^{[5]} = \begin{bmatrix} 0.87 & -1.04 & 0.00 \\ 0.46 & -0.69 & 0.00 \\ 0.00 & -0.29 & 0.00 \end{bmatrix}, \quad \hat{\Sigma}_a^{[5]} = \begin{bmatrix} 2.10 & & \\ 2.19 & 2.42 & \\ 2.07 & 2.27 & 2.66 \end{bmatrix} \times 10^{-3} \qquad [6.5]$$

Model [6]: $(I - \Phi_1^{[6]} B)\mathbf{z}_t = \mathbf{a}_t$

$$\hat{\Phi}_1^{[6]} = \begin{bmatrix} -0.87 & 1.02 & 0.00 \\ -0.44 & 0.63 & 0.00 \\ 0.00 & 0.25 & 0.00 \end{bmatrix}, \quad \hat{\Sigma}_a^{[6]} = \begin{bmatrix} 2.05 & & \\ 2.16 & 2.42 & \\ 2.07 & 2.29 & 2.69 \end{bmatrix} \times 10^{-3} \qquad [6.6]$$

Some fitting results for these are summarised in Table 6.5 and the residuals from model [5] are plotted in Figures 6.3d–f. There are no particularly large outliers and the residuals appear to be uncorrelated, so that the model fits reasonably well. Model [6] produces very similar residuals.

6.2.3 T&T

In their analysis of this data, T&T chose to fit a VARMA(1,1) model to a linear transformation of the data. To see why, we present the root table $R(m,h,j)$ (with $h=m$) in Table 6.6(a) for the raw (logged but undifferenced) data. This suggests that either a VARMA(1,1) model consisting of SCMs (1,0), (1,0) and (1,1) or a VAR(2) model consisting of SCMs (1,0), (1,0) and (2,0) would be appropriate. T&T only chose to estimate the VARMA(1,1) model, whilst Tiao and Tsay (1985) consider the VAR(2) model. We have re–estimated both of these by exact maximum likelihood and transformed them back in terms of the original data (all mean–corrected).

Model [7]: $(I - \Phi_1 B)\mathbf{z}_t = (I - \Theta_1 B)\mathbf{a}_t$, $T^{[7]} = \begin{bmatrix} -0.40 & 0.83 & -0.40 \\ 0.61 & -0.51 & -0.60 \\ 0.55 & 0.83 & -0.06 \end{bmatrix}$, [6.7]

$$\hat{\Phi}_1^{[7]} = \begin{bmatrix} 1.23 & -0.47 & 0.19 \\ 0.28 & 0.46 & 0.20 \\ 0.25 & -0.46 & 1.15 \end{bmatrix}, \hat{\Theta}_1^{[7]} = \begin{bmatrix} 1.17 & -1.04 & -0.08 \\ 0.79 & -0.69 & -0.05 \\ 0.40 & -0.35 & -0.03 \end{bmatrix}, \hat{\Sigma}_a^{[7]} = \begin{bmatrix} 2.02 & & \\ 2.10 & 2.32 & \\ 2.03 & 2.22 & 2.63 \end{bmatrix} \times 10^{-3}$$

Model [8]: $(I-\Phi_1 B-\Phi_2 B^2)\mathbf{z}_t=\mathbf{a}_t$, $T^{[8]}=\begin{bmatrix}0.44 & -0.83 & 0.34\\ 0.47 & -0.19 & -0.86\\ 0.84 & -0.27 & -0.17\end{bmatrix}$, [6.8]

$$\hat{\Phi}_1^{[8]}=\begin{bmatrix}-0.11 & 0.74 & 0.28\\ -0.66 & 1.29 & 0.28\\ -0.32 & 0.04 & 1.21\end{bmatrix},\hat{\Phi}_2^{[8]}=\begin{bmatrix}1.25 & -1.15 & -0.05\\ 0.88 & -0.81 & -0.04\\ 0.54 & -0.49 & -0.02\end{bmatrix},\hat{\Sigma}_a^{[8]}=\begin{bmatrix}1.90 & & \\ 1.98 & 2.21 & \\ 1.94 & 2.13 & 2.57\end{bmatrix}\times 10^{-3}$$

The residuals from model [6.7] are plotted in Figures 6.3g–i and those from model [6.8] are similar. The residual series all appear to be satisfactory and suggest that the models are both adequate.

From the CI analysis we are led to consider differencing each of the series and, applying T&T to $\nabla\mathbf{z}_t$ gives the root table of Table 6.6b. This suggests either a VMA(1) or VAR(1) model for the differenced data consisting of 2 SCM(0,0)s and either an SCM(0,1) or SCM(1,0). We can also estimate these models and transform them back to obtain the following.

Model [9]: $\nabla\mathbf{z}_t=(I-\Theta_1 B)\mathbf{a}_t$, $T^{[9]}=\begin{bmatrix}-0.39 & 0.82 & -0.41\\ 0.63 & -0.54 & -0.57\\ 0.77 & 0.54 & 0.34\end{bmatrix}$

$$\hat{\Theta}_1^{[9]}=\begin{bmatrix}1.22 & -1.07 & -0.11\\ 0.85 & -0.75 & -0.08\\ 0.54 & -0.47 & -0.05\end{bmatrix},\ \hat{\Sigma}_a^{[9]}=\begin{bmatrix}2.18 & & \\ 2.29 & 2.54 & \\ 2.19 & 2.41 & 2.83\end{bmatrix}\times 10^{-3} \qquad [6.9]$$

Model [10]: $(I-\Phi_1 B)\nabla\mathbf{z}_t=\mathbf{a}_t$, $T^{[10]}=\begin{bmatrix}-0.43 & 0.83 & -0.36\\ -0.52 & 0.21 & 0.83\\ 0.90 & 0.23 & -0.38\end{bmatrix}$

$$\hat{\Phi}_1^{[10]}=\begin{bmatrix}-1.32 & 1.20 & 0.09\\ -0.95 & 0.86 & 0.07\\ -0.60 & 0.54 & 0.04\end{bmatrix},\ \hat{\Sigma}_a^{[10]}=\begin{bmatrix}2.11 & & \\ 2.22 & 2.49 & \\ 2.15 & 2.38 & 2.80\end{bmatrix}\times 10^{-3} \qquad [6.10]$$

These models appear to fit reasonably well with residuals comparable with those of models [6.5] and [6.6].

Application of the method of T89b to the Flour price data (differenced or undifferenced) results in similar SCM and model orders to those suggested by T&T, so we will not present the analysis. As discussed in §5.5.4, both procedures use identical constructed covariance matrices and T89b should only prove to be

quicker when higher SCM orders are necessary.

6.2.4 Model fitting

The values of AIC have been calculated for each of the models and are presented in Table 6.5. As discussed in §6.1.7, we have not taken into account the "parameters" in the matrix T for the models built using T&T's method. The values are also presented for AIC in the form of [3.11] (though this should not be used for other than pure VAR models) and in this case the differences are not constant, even between the pure VAR models, although the ordering remains the same.

6.2.5 Forecasting

In order to compare the forecasting performance of the models, we refitted them to the first 89 observations (to December 1979) and used these to forecast the final years' data (from January to November 1980, 11 observations). All models were fitted to the mean–corrected data and the forecasts calculated by adding back the mean. For the transformed models we used the untransformed form of the model to forecast (although the results are the same for an invertible transformation matrix (see §3.5.1(c)) as all the Ts are). We plot the forecasts of the first component series (Buffalo) for each of the six models in Figure 6.4. The sum of squared forecasting errors over the 3 components is tabulated in Table 6.7a for each model and forecasting horizon and estimates of the standard errors of the forecast errors for the first component series are given in Table 6.7b (see §2.5.2 and §3.3.2).

6.2.6 Comparison

These results suggest that model [8] is the best fitting and provides the best out–of–sample forecasts at longer horizons suggesting that perhaps it captures more of the long–term influences in the data. However, model [7] produces the best forecasts at shorter horizons (perhaps helped by the drop in April 1980 – a comparison using only the last 6 months, say would probably reach different conclusions) and this model only fits the data worse than model [8] by about 2 parameters. Both models produce more precise forecasts than the others (as measured by the estimates of the standard errors of the forecast errors in Table 6.7b). Notice that for the non–stationary model [5], these standard errors grow rapidly for longer horizons. For the differenced data, models [9] and [10] fit the data at least as well as models [5] and [6] and produce the better forecasts.

In conclusion, for this dataset, the model specification procedures appear to be able to specify better models (in terms of both fit and out–of–sample forecasting performance) than the more usual model identification tools of Chapter 2.

Relatively little effort was required to specify these models using Tsay's FORTRAN programs and the estimations were better conditioned (and quicker to converge – although none of the estimations took more than a few seconds in this case) than those for the untransformed data ([5] and [6]). It also appears that better fitting models producing superior forecasts can be obtained by working with the undifferenced data (models [7] and [8]), although with n.s. data this will not always be possible and the use of a model specification procedure then becomes necessary. The effect of n.s. on the modelling clearly needs further investigation.

Table 6.4a Cross correlations.

Lag	ccf	$2/\sqrt{N}$ significance	Accurate significance
1	0.14 0.16 0.20 0.22 0.22 0.24 0.20 0.19 0.18	. . . + + + . . .	. . . + + + . . .
2	0.04 0.01 0.01 0.00 −0.02 −0.01 0.01 −0.00 0.00	. . . .	. . . .
3	−0.03 0.01 0.05 −0.02 0.01 0.04 0.03 0.07 0.06	. . . .	. . . .
4	0.02 0.04 0.05 −0.00 0.01 0.04 0.02 0.03 0.05	. . . .	. . . .
5	0.02 0.03 −0.01 −0.00 0.01 −0.03 0.04 0.04 −0.00	. . . .	. . . .

Collected significances

$2/\sqrt{N}$			Accurate		
.....	.....	.....	.....	.....	.....
+....	+....	+....	+....	+....	+....
.....	.....	.....	.....	.....	.....

Table 6.4b Partial autoregressions.

Lag	parm	Significance	M	p-value
1	−1.25 1.37 −0.01 −0.84 1.03 −0.04 −0.48 0.83 −0.15	− + . − + . − + .	43.48	0
2	0.31 −0.40 0.08 0.40 −0.51 0.09 0.44 −0.53 0.07	+ − . + − . + − .	1.71	1
3	0.07 −0.23 0.19 0.16 −0.35 0.26 0.23 −0.24 0.10	. − . . − + + − .	−0.56	1
4	0.24 −0.30 0.02 0.51 −0.55 0.03 0.53 −0.47 −0.02	+ − . + − . + − .	6.63	0.68

5	$\begin{bmatrix} 0.37 & -0.58 & 0.25 \\ 0.45 & -0.59 & 0.20 \\ 0.56 & -0.66 & 0.12 \end{bmatrix}$	$\begin{bmatrix} + & - & + \\ + & - & . \\ + & - & . \end{bmatrix}$	−6.6	1

Table 6.4c Partial lag correlations.

Lag	plcm	Significance	X	p-value
1	$\begin{bmatrix} 0.14 & 0.16 & 0.20 \\ 0.22 & 0.22 & 0.24 \\ 0.20 & 0.19 & 0.18 \end{bmatrix}$	$\begin{bmatrix} . & . & . \\ + & + & + \\ . & . & . \end{bmatrix}$	34.63	0
2	$\begin{bmatrix} -0.02 & -0.04 & -0.04 \\ -0.04 & -0.06 & -0.06 \\ -0.02 & -0.04 & -0.03 \end{bmatrix}$	$\begin{bmatrix} . & . & . \\ . & . & . \\ . & . & . \end{bmatrix}$	1.52	1
3	$\begin{bmatrix} 0.02 & 0.05 & 0.07 \\ 0.02 & 0.04 & 0.07 \\ 0.06 & 0.09 & 0.08 \end{bmatrix}$	$\begin{bmatrix} . & . & . \\ . & . & . \\ . & . & . \end{bmatrix}$	3.3	0.95
4	$\begin{bmatrix} -0.05 & -0.02 & 0.02 \\ -0.07 & -0.05 & -0.00 \\ -0.05 & -0.03 & -0.00 \end{bmatrix}$	$\begin{bmatrix} . & . & . \\ . & . & . \\ . & . & . \end{bmatrix}$	1.42	1
5	$\begin{bmatrix} 0.01 & 0.03 & -0.00 \\ -0.00 & 0.01 & -0.03 \\ 0.06 & 0.06 & -0.00 \end{bmatrix}$	$\begin{bmatrix} . & . & . \\ . & . & . \\ . & . & . \end{bmatrix}$	0.84	1

Table 6.5 Flour price data model fitting information.

Model	LL	Npar	$AIC_{[3.9]}$	(min=0)	$AIC_{[3.11]}$	(min=0)
[5]	697.3	5	−1384.5	14.7	−2227.5	30.7
[6]	698.6	5	−1387.2	12.1	−2230.8	27.4
[7]	705.4	8	−1394.8	4.5	−2253.7	4.5
[8]	708.6	9	−1399.3	0.0	−2258.1	0.0
[9]	694.9	3	−1383.8	15.4	−2227.1	31.1
[10]	696.9	3	−1387.8	11.4	−2231.5	26.6

Table 6.6 Root tables for Flour price data.

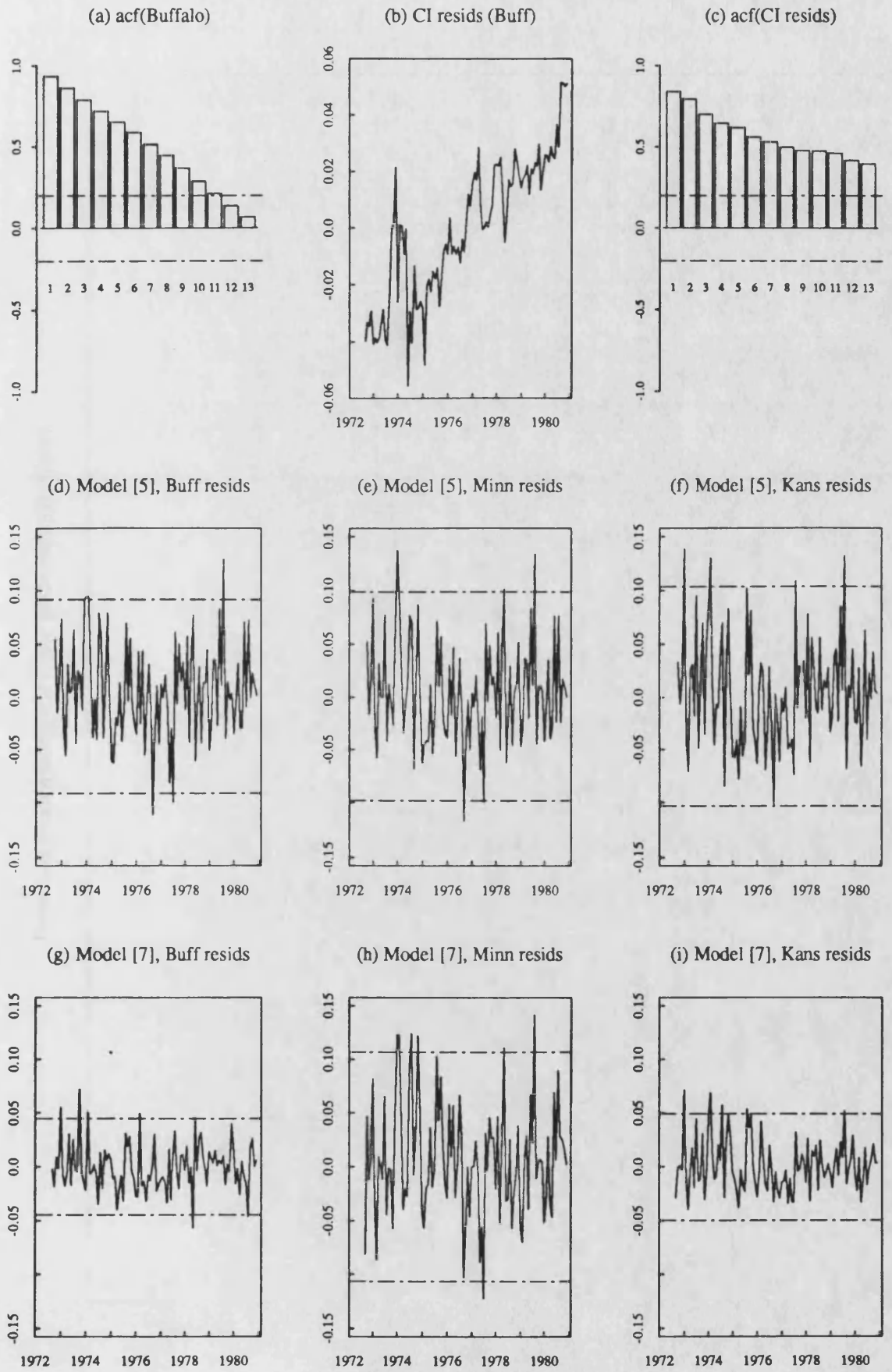
(a) Raw data

$m\backslash j$	0	1	2	3	4
0	0	0	0	0	0
1	2	3	3	3	3
2	3	6	6	6	6
3	3	6	9	9	9
4	3	6	9	12	12

(b) Differenced data

$m\backslash j$	0	1	2	3	4
0	2	3	3	3	3
1	3	6	6	6	6
2	3	6	9	9	9
3	3	6	9	12	12
4	3	6	9	12	15

Figure 6.3


(a) acf(Buffalo)
(b) CI resids (Buff)
(c) acf(CI resids)
(d) Model [5], Buff resids
(e) Model [5], Minn resids
(f) Model [5], Kans resids
(g) Model [7], Buff resids
(h) Model [7], Minn resids
(i) Model [7], Kans resids
1 2 3 4 5 6 7 8 9 10 11 12 13
1972 1974 1976 1978 1980

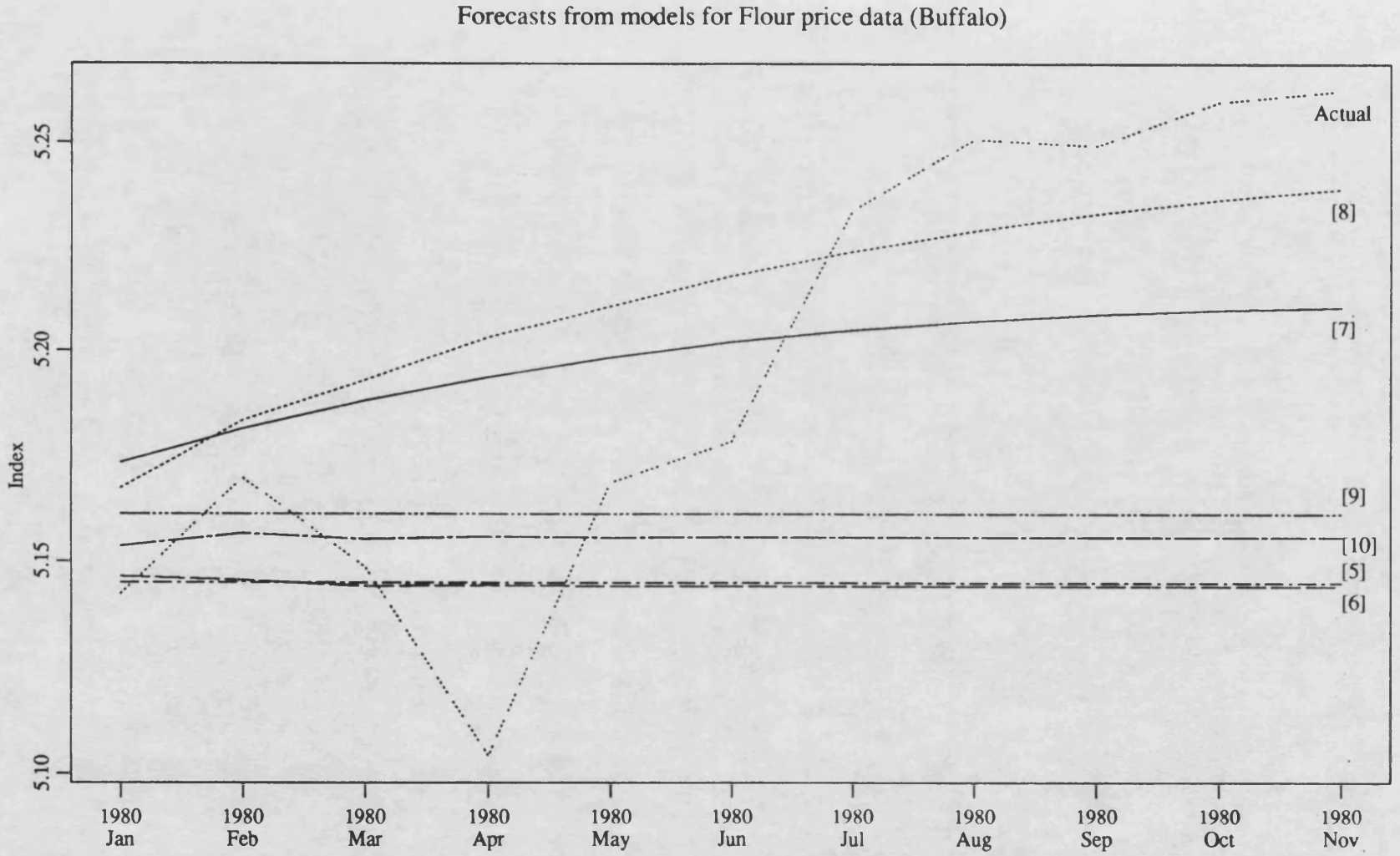

Forecasts from models for Flour price data (Buffalo)
Index
5.25
5.20
5.15
5.10
Actual
[8]
[7]
[9]
[10]
[5]
[6]
1980 Jan
1980 Feb
1980 Mar
1980 Apr
1980 May
1980 Jun
1980 Jul
1980 Aug
1980 Sep
1980 Oct
1980 Nov

Figure 6.4

Table 6.7a Sum of squared forecasting errors $\times 10^4$.

Horizon	Model					
	[5]	[6]	[7]	[8]	[9]	[10]
1	6	8	42	38	22	15
2	13	13	10	16	3	5
3	29	27	116	149	51	45
4	115	111	334	419	175	159
5	21	21	75	129	23	22
6	40	39	110	187	45	44
7	158	160	71	106	120	132
8	174	177	52	79	126	140
9	176	180	21	32	118	134
10	226	231	30	23	158	176
11	241	247	32	19	170	189
Total	1198	1212	891	1198	1011	1062

Table 6.7b Forecast error standard errors $\times 10^4$ (Buffalo).

Horizon	Model					
	[5]	[6]	[7]	[8]	[9]	[10]
1	470	465	471	447	484	474
2	762	745	640	628	695	697
3	1095	963	762	750	870	851
4	1503	1145	860	853	1023	986
5	2017	1302	941	943	1159	1102
6	2675	1443	1012	1026	1282	1208
7	3524	1571	1077	1106	1396	1305
8	4623	1689	1136	1183	1501	1396
9	6050	1800	1192	1258	1599	1481
10	7908	1904	1246	1332	1692	1561
11	10328	2003	1297	1404	1780	1637

6.3 UK Pig production data

The UK pig production data is described in Appendix A.3 and plotted and tabulated there. It has some quarterly seasonality present which we correct for using the seasonal adjustment method of Cleveland and Devlin (1982) implemented by S–Plus – see e.g Figure 6.5 which displays the original series 1 (Gilts), together with the estimated (additive) seasonal component (notice the small scale of this component) and seasonally adjusted data. We work with the normalised data to make the scales comparable. The correlations of the adjusted data are shown in Tables 6.8 suggesting that perhaps a VAR(1) model would be appropriate (individually, series 4 and 5 may possibly be n.s., but are not found to be CI. We will analyse the undifferenced series). The results from estimating a VAR(1) model, iteratively deleting parameters are given below.

Model [11]: $(I-\Phi_1^{[11]}B)\mathbf{z}_t=\mathbf{a}_t$

$$\hat{\Phi}_1^{[11]}=\begin{bmatrix} 0.63 & 0.49 & 0.00 & 0.00 & 0.00 \\ -0.35 & 0.81 & -0.17 & 0.00 & 0.00 \\ 0.00 & -0.16 & 0.92 & -0.33 & 0.42 \\ 0.00 & -0.16 & 0.00 & 0.60 & 0.35 \\ 0.20 & 0.14 & -0.17 & 0.24 & 0.61 \end{bmatrix}, \quad \hat{\Sigma}_a^{[11]}=\begin{bmatrix} 9.24 & & & & \\ 2.57 & 30.99 & & & \\ -1.53 & -15.47 & 24.13 & & \\ 1.74 & -6.84 & 1.79 & 7.36 & \\ 4.22 & -2.06 & -0.62 & 3.01 & 6.12 \end{bmatrix} \times 10^{-2}$$

Application of the method T&T results in an overall VARMA(1,1) model (consisting of SCMs of orders (0,1), 3×(1,0) and (1,1)) for some transformation of the data, which we also estimate. The results are given below, transformed back in terms of the original data.

Model [12]: $(I-\Phi_1^{[12]}B)\mathbf{z}_t=(I-\Theta_1^{[12]}B)\mathbf{a}_t$,

$$T=\begin{bmatrix} -0.47 & 0.11 & 0.20 & -0.57 & 0.63 \\ 0.42 & -0.36 & -0.04 & 0.38 & 0.74 \\ -0.57 & 0.58 & 0.23 & 0.52 & -0.12 \\ 0.82 & -0.21 & 0.10 & -0.10 & 0.52 \\ -0.25 & -0.51 & 0.78 & 0.13 & -0.22 \end{bmatrix}, \hat{\Phi}_1^{[12]}=\begin{bmatrix} 0.17 & 0.52 & 0.11 & -0.51 & 0.61 \\ -1.88 & 1.01 & 0.30 & -1.76 & 2.05 \\ 2.49 & -0.38 & 0.13 & 2.70 & -2.99 \\ 0.07 & -0.20 & -0.02 & 0.64 & 0.26 \\ -0.28 & 0.15 & -0.03 & -0.34 & 1.27 \end{bmatrix},$$

$$\hat{\Theta}_1^{[12]}=\begin{bmatrix} -0.42 & 0.03 & 0.09 & -0.42 & 0.38 \\ -1.83 & 0.00 & 0.40 & -1.87 & 2.14 \\ 2.87 & -0.03 & -0.63 & 2.93 & -3.27 \\ 0.20 & 0.03 & -0.05 & 0.21 & -0.35 \\ -0.60 & -0.04 & 0.14 & -0.62 & 0.83 \end{bmatrix}, \hat{\Sigma}_a^{[12]}=\begin{bmatrix} 8.95 & & & & \\ 1.18 & 25.40 & & & \\ -0.69 & -7.09 & 15.23 & & \\ 2.27 & -6.88 & 0.12 & 7.96 & \\ 3.60 & -2.48 & 1.09 & 3.16 & 5.06 \end{bmatrix} \times 10^{-2}$$

Both of these models give residual series which are satisfactory (uncorrelated and with no large outliers) – see Figures 6.6 and 6.7 which plot the series together with $\pm 2\sqrt{[\hat{\Sigma}_a]_{ii}}$ limits. These suggest that both models fit the data quite well. Model [12] can be seen to fit the data better than model [11] by an amount equivalent to about 7 parameters (see the AIC values in Table 6.9). The short amount of data available prevent us from making a detailed comparison of the forecasting performance of these two models – the results of §6.1 and §6.2 suggest that differences between the models would not become apparent until we were forecasting at longer horizons.

For this dataset, the model specification procedure of T&T appears to have helped us to specify a better fitting model than would be arrived at with the more usual approach (discussed in Chapter 2). We have not applied T89b to this data since the structure appears to be simple and there would be no gain to be made with the more efficient search path. Notice that the correlation matrices are becoming very difficult to use for order determination and a model specification method is essential

Table 6.8a Cross correlations.

Lag	ccf					$2/\sqrt{N}$ significance					Accurate significance				
	0.85	0.14	−0.22	−0.01	0.65	+	.	.	.	+	+	.	.	.	+
	0.79	0.76	−0.67	−0.48	0.23	+	+	−	−	.	+	+	−	−	.
1	−0.68	−0.60	0.83	0.31	−0.36	−	−	+	+	−	−	−	+	.	.
	−0.33	−0.38	0.37	0.90	0.46	−	−	+	+	+	.	−	.	+	+
	0.37	−0.14	0.09	0.70	0.90	+	.	.	+	+	.	.	.	+	+
	0.58	−0.22	0.09	0.19	0.61	+	.	.	.	+	+	.	.	.	+
	0.87	0.45	−0.45	−0.33	0.40	+	+	−	−	+	+	+	−	.	.
2	−0.73	−0.36	0.57	0.14	−0.54	−	−	+	.	−	−	.	+	.	−
	−0.38	−0.21	0.30	0.75	0.33	−	.	+	+	+	.	.	.	+	.
	0.16	−0.26	0.26	0.74	0.76	.	.	.	+	+	.	.	.	+	+

	0.25	−0.42	0.26	0.34	0.50		.	−	.	+	+		.	−	.	.	+
	0.75	0.09	−0.14	−0.12	0.46		+	.	.	.	+		+	.	.	.	+
3	−0.60	−0.13	0.28	−0.04	−0.58		−	.	.	.	−		−	.	.	.	−
	−0.33	−0.03	0.16	0.59	0.22		−	.	.	+	.		.	.	.	+	.
	−0.03	−0.26	0.29	0.73	0.60		.	.	+	+	+		.	.	.	+	+
	−0.07	−0.46	0.31	0.40	0.32		.	−	+	+	+		.	−	.	.	.
	0.51	−0.26	0.13	0.05	0.45		+	.	.	.	+		+	.	.	.	+
4	−0.42	0.10	0.01	−0.22	−0.57		−	.	.	.	−		.	.	.	.	−
	−0.23	0.10	0.05	0.42	0.17		.	.	.	+	.		.	.	.	.	.
	−0.16	−0.18	0.26	0.64	0.44		.	.	.	+	+		.	.	.	+	.
	−0.28	−0.34	0.24	0.36	0.16		.	−	.	+	.		.	.	.	.	.
	0.19	−0.45	0.34	0.21	0.34		.	−	+	.	+		.	−	.	.	.
5	−0.24	0.20	−0.14	−0.37	−0.51		.	.	.	−	−		.	.	.	.	−
	−0.13	0.14	−0.01	0.28	0.15		.	.	.	.	.		.	.	.	.	.
	−0.20	−0.06	0.15	0.51	0.30		.	.	.	+	+		.	.	.	.	.

Collected significances

$2/\sqrt{N}$

++...	..−−−	...+.	..+++	++++.
++++.	++..−	−−..+	−−...	.++++
−−−−.	−−...	++...	+...−	−−−−−
−−−..	−....	++...	++++.	++...
+....	.....	..+..	+++++	+++++

Accurate

++...	..−−.	.....	.....	+++..
++++.	++..−	−−...	−....	..++.
−−−..	−....	++...	.....	.−−−−
.....	−....	.....	+++..	+....
.....	.....	.....	++++.	+++..

Table 6.8b Partial autoregressions.

Lag	parm					Significance					*M*	p-value
	0.67	0.49	−0.06	0.14	−0.11	+	+	.	.	.		
	−0.42	0.83	−0.15	−0.05	0.12	−	+	.	.	.		
1	−0.03	−0.19	0.86	−0.35	0.44	.	.	+	−	+	334.08	0
	0.02	−0.16	−0.00	0.64	0.29	.	.	.	+	.		
	0.23	0.13	−0.19	0.30	0.55	.	.	.	+	+		
	0.08	0.09	−0.05	−0.11	−0.08	.	.	.	.	.		
	−0.74	−0.14	0.26	0.01	0.11	−	.	.	.	.		
2	0.57	0.32	−0.08	0.22	−0.23	+	+	.	.	.	−135.43	1
	0.20	0.06	−0.24	0.07	−0.12	.	.	.	.	.		
	−0.02	0.06	−0.23	−0.02	0.13	.	.	.	.	.		
	0.23	0.02	0.20	−0.12	−0.16	.	.	.	.	.		
	0.08	0.01	−0.10	−0.01	0.18	.	.	.	.	.		
3	−0.28	−0.01	−0.03	−0.17	−0.37	.	.	.	.	−	−57.47	1
	−0.10	0.04	0.07	−0.10	0.10	.	.	.	.	.		
	0.18	−0.05	0.12	−0.13	−0.07	.	.	.	.	.		

4	$\begin{bmatrix} -0.05 & 0.06 & 0.00 & -0.22 & 0.07 \\ -0.55 & -0.28 & 0.01 & -0.43 & 0.64 \\ 0.08 & 0.13 & -0.06 & 0.07 & -0.07 \\ -0.08 & -0.27 & -0.10 & -0.01 & -0.27 \\ -0.00 & 0.00 & -0.07 & 0.02 & -0.01 \end{bmatrix}$	$\begin{bmatrix} . & . & . & . & . \\ - & . & . & - & + \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix}$	−14.71	1
5	$\begin{bmatrix} 0.13 & 0.10 & -0.14 & -0.29 & -0.14 \\ -0.30 & -0.46 & 0.12 & -0.50 & 0.31 \\ 0.08 & 0.36 & 0.07 & 0.10 & -0.29 \\ -0.00 & 0.07 & -0.26 & 0.04 & 0.00 \\ 0.17 & 0.09 & -0.12 & 0.01 & -0.28 \end{bmatrix}$	$\begin{bmatrix} . & . & . & - & . \\ - & - & . & - & + \\ . & + & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix}$	−16.66	1

Table 6.8c Partial lag correlations.

Lag	plcm	Significance	X	p-value
1	$\begin{bmatrix} 0.85 & 0.14 & -0.22 & -0.01 & 0.65 \\ 0.79 & 0.76 & -0.67 & -0.48 & 0.23 \\ -0.68 & -0.60 & 0.83 & 0.31 & -0.36 \\ -0.33 & -0.38 & 0.37 & 0.90 & 0.46 \\ 0.37 & -0.14 & 0.09 & 0.70 & 0.90 \end{bmatrix}$	$\begin{bmatrix} + & . & . & . & + \\ + & + & - & - & . \\ - & - & + & + & - \\ - & - & + & + & + \\ + & . & . & + & + \end{bmatrix}$	371	0
2	$\begin{bmatrix} -0.02 & -0.39 & 0.25 & 0.07 & 0.00 \\ 0.16 & -0.01 & 0.17 & 0.08 & 0.11 \\ -0.09 & 0.14 & -0.05 & -0.23 & -0.30 \\ -0.20 & 0.04 & 0.05 & -0.01 & -0.00 \\ -0.13 & -0.23 & 0.12 & 0.03 & 0.09 \end{bmatrix}$	$\begin{bmatrix} . & - & . & . & . \\ . & . & . & . & . \\ . & . & . & . & - \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix}$	28.83	0.27
3	$\begin{bmatrix} 0.17 & 0.12 & -0.43 & -0.04 & 0.17 \\ -0.06 & -0.03 & 0.18 & 0.05 & -0.11 \\ 0.30 & -0.06 & -0.09 & 0.04 & 0.17 \\ -0.17 & 0.05 & -0.29 & -0.05 & -0.12 \\ -0.07 & 0.15 & -0.49 & -0.02 & 0.01 \end{bmatrix}$	$\begin{bmatrix} . & . & - & . & . \\ . & . & . & . & . \\ + & . & . & . & . \\ . & . & - & . & . \\ . & . & - & . & . \end{bmatrix}$	40.76	0.02
4	$\begin{bmatrix} -0.04 & -0.04 & -0.01 & -0.13 & -0.02 \\ 0.13 & -0.17 & 0.12 & -0.19 & 0.02 \\ -0.06 & -0.02 & -0.07 & -0.05 & -0.08 \\ -0.25 & -0.09 & 0.01 & -0.07 & 0.00 \\ -0.09 & 0.12 & -0.02 & -0.20 & -0.01 \end{bmatrix}$	$\begin{bmatrix} . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix}$	13.11	0.98
5	$\begin{bmatrix} -0.01 & 0.09 & -0.21 & -0.05 & -0.05 \\ 0.24 & -0.24 & 0.32 & 0.12 & 0.14 \\ -0.27 & 0.10 & 0.01 & -0.28 & -0.15 \\ -0.46 & -0.14 & -0.07 & -0.03 & -0.12 \\ -0.24 & 0.08 & -0.26 & -0.01 & -0.18 \end{bmatrix}$	$\begin{bmatrix} . & . & . & . & . \\ . & . & + & . & . \\ . & . & . & . & . \\ - & . & . & . & . \\ . & . & . & . & . \end{bmatrix}$	**44.03**	**0.01**

Table 6.9 Model fitting information for Pig data.

Model	LL	Npar	$AIC_{[3.9]}$	(min=0)	$AIC_{[3.11]}$	(min=0)
[11] (1,0)	-62.5	17	159.0	14.9	-529.8	17.1
[12] (1,1)	-49.0	23	144.1	0	-546.9	0

Figure 6.5 Seasonal adjustment of Gilts

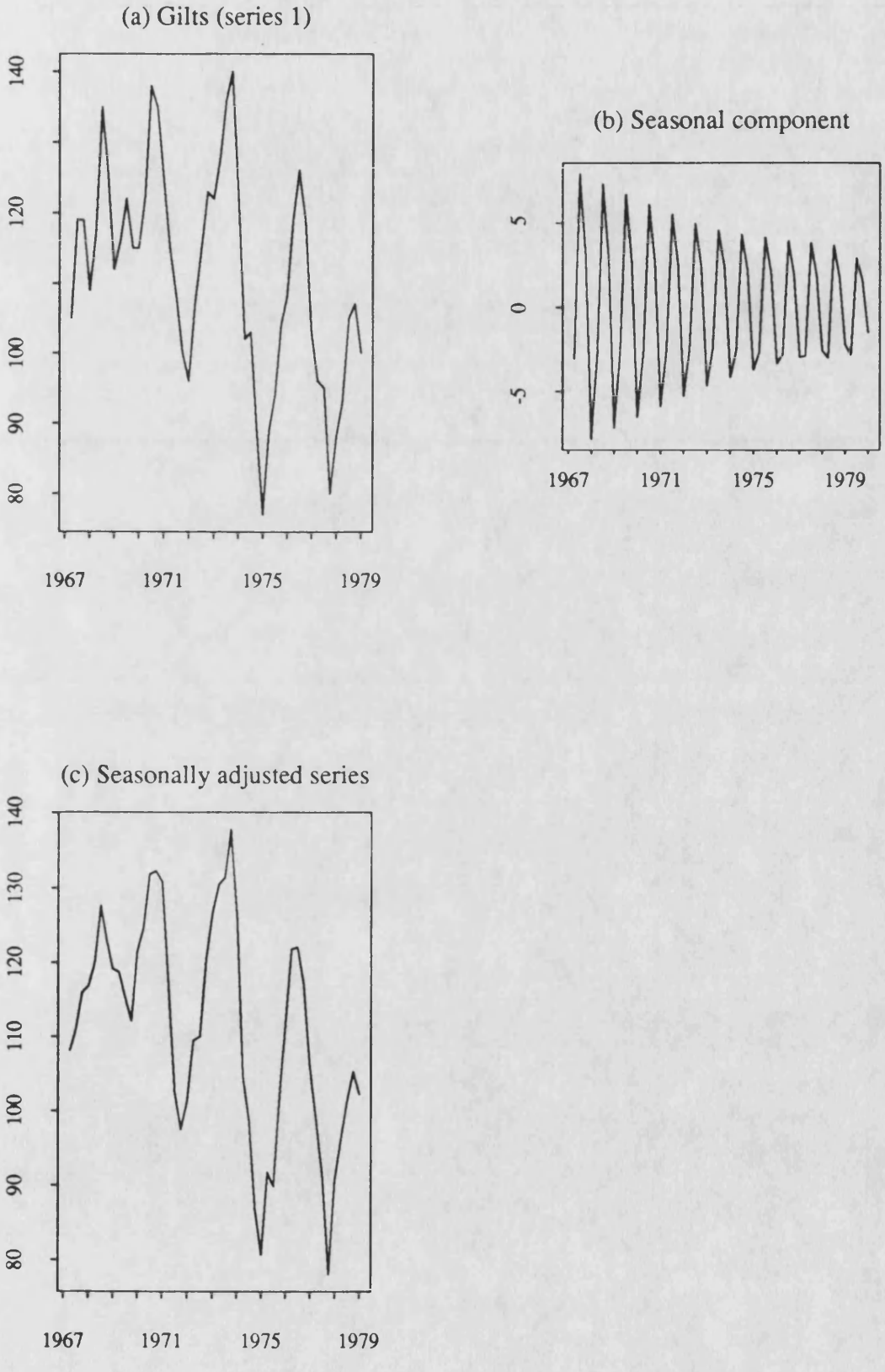
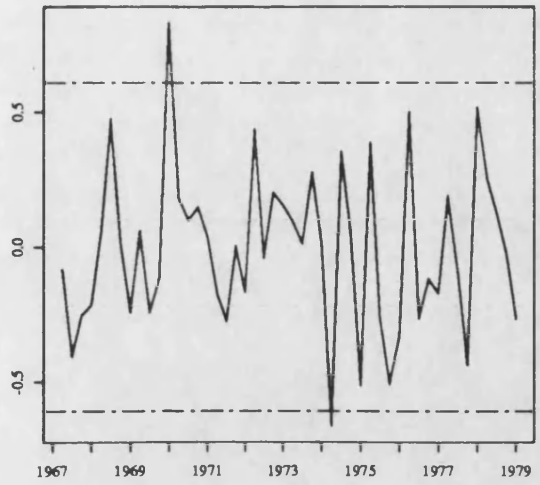
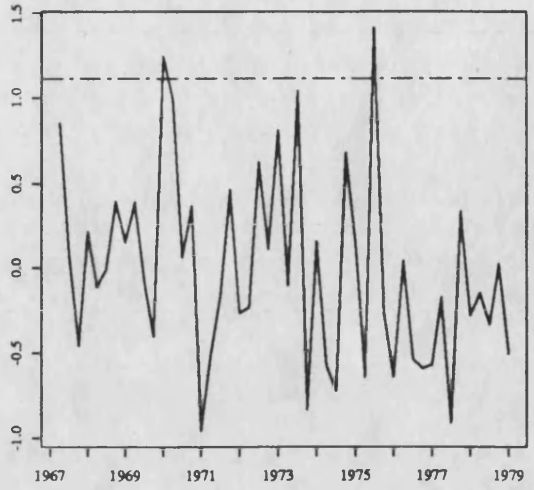


Figure 6.6 Residuals from model [11]

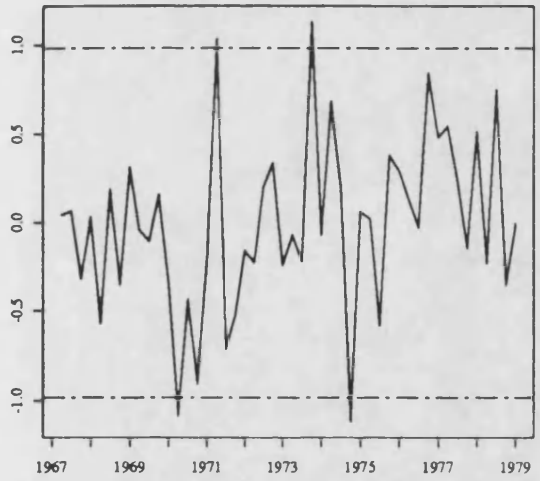
(a) Series 1



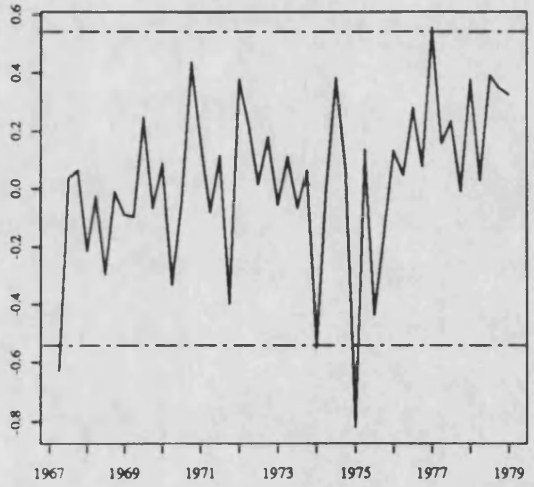
(b) Series 2



(c) Series 3



(d) Series 4



(e) Series 5

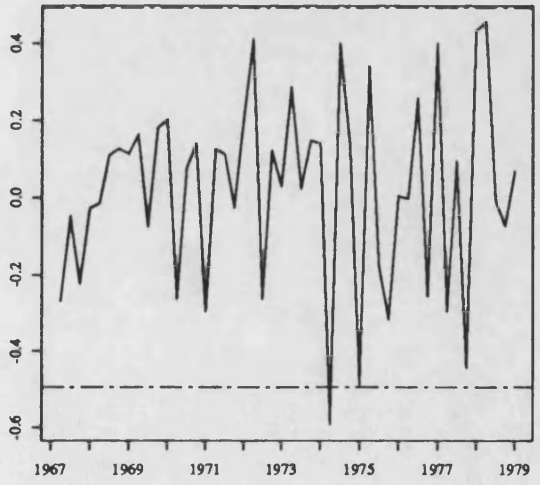
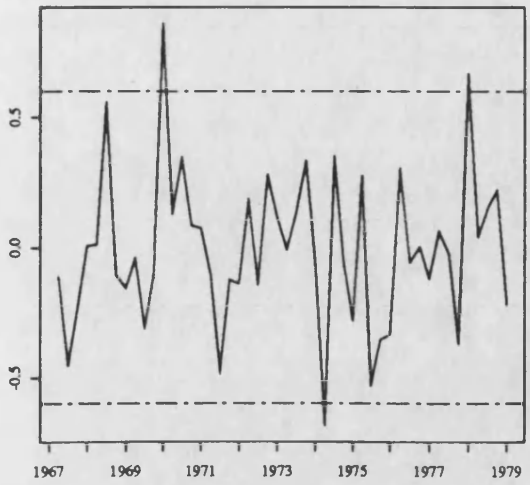
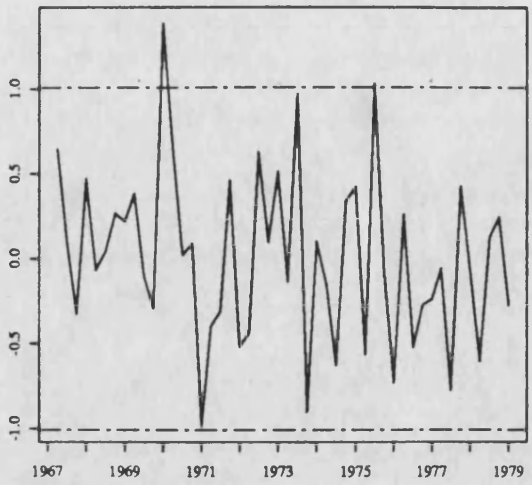


Figure 6.7 Residuals from model [12]

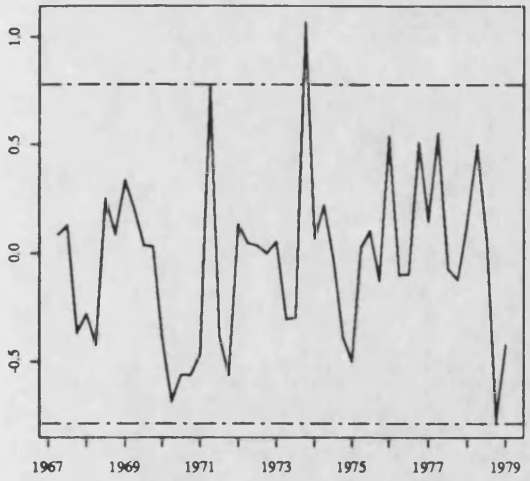
(a) Series 1



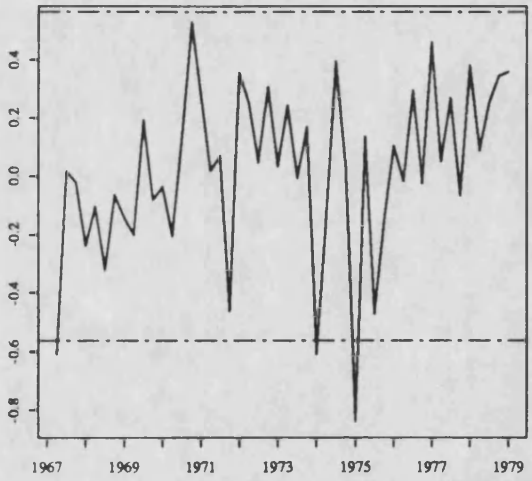
(b) Series 2



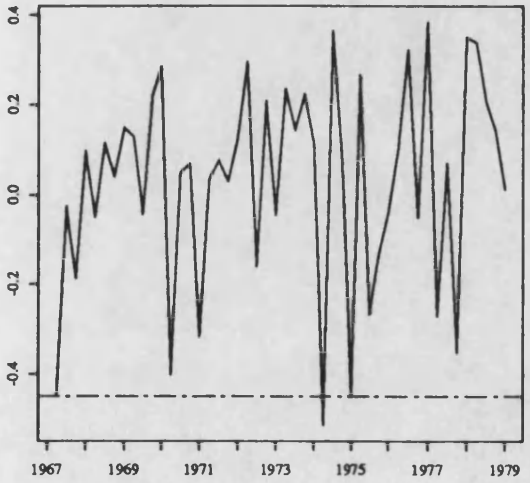
(c) Series 3



(d) Series 4



(e) Series 5



6.4 US Hog data

This dataset is described in Appendix A.4 and the original data are tabulated in Table A.4 and plotted in Figure A.4. It was seen in §4.5 that the series are n.s. and CI with 3 linearly independent CI vectors. The definitions of the variables given in §A.4 show that the series are already logged, but as mentioned there, T&T chose to work with the log of this data (i.e. log(log)), to reduce the scale. Application of the method of T&T to $\mathbf{y}_t=\ln(\mathbf{z}_t)$ in their paper suggested a VARMA(2,1) model for some linear transformation of $\mathbf{y}_t$, consisting of SCMs of orders (0,0), (0,1), (1,0), (1,1) and (2,0) (we will call this model [13]). T&T state that two of the transformed variables are stationary (whereas the original were each n.s.) suggesting that the analysis has found two CI vectors (see §4.7.3 and §5.2.12). This VARMA structure contains at most 29 parameters (using the elimination rule of §5.2.3). For the "raw" data $\mathbf{z}_t$ (as described in §A.4, but normalised), application of T&T suggests a VARMA(1,1) model (model [14]), for a different linear transformation, consisting of SCMs of orders (0,0), (0,1) and 3 SCM(1,1)s and containing at most 32 parameters. These discrepancies bring us to question the further transformation of this data (to $\mathbf{y}_t$) and it may be fruitful to consider the optimal choice of a (power) transformation parameter λ (see §3.1.1(b)). However (partly because of the nature and shortcomings of this dataset, discussed in §A.4) we will not pursue this point, but instead consider the CI found in §4.5 and its effect on the modelling.

6.4.1 Co–integrated models

The raw data are shown to be n.s. and CI, so that the transformed and selectively differenced data, $D(B)T\mathbf{z}_t$ is "minimally" stationary (stationary without being overdifferenced) where

$$D(B)=(I-D_1B)=\begin{bmatrix} I_3 & \\ & \nabla I_2 \end{bmatrix} \text{ and } T=\begin{bmatrix} 1.00 & 0.34 & -0.42 & -0.20 & -0.32 \\ -1.07 & -0.74 & 1.05 & 1.00 & 0.00 \\ -0.92 & -0.71 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

(see §4.5.1). Application of T&T to $SD(B)T\mathbf{z}_t$ (where S is a normalising matrix – see §3.5.2(a)) suggests that a VARMA(1,1) model (model [15]) consisting of 2×SCM(0,0)s, 2×SCM(0,1)s and an SCM(1,1) may be satisfactory for a transformation P of this (we use "P" now for the transformation produced by T&T since we already have "T" as the transformation from the CI analysis). We can estimate this model to give us

model [15] $\quad (I-\Phi_1B)PSD(B)T\mathbf{z}_t=(I-\Theta_1B)\mathbf{b}_t,$

and this can be transformed back in terms of $\mathbf{z}_t$ to give

$$(I-\Phi_1 B)PS(I-D_1 B)T\mathbf{z}_t=(I-\Theta_1 B)\mathbf{b}_t$$

$$(I-\Phi_1 B)(PST-(PSD_1 T)B)\mathbf{z}_t=(I-\Theta_1 B)\mathbf{b}_t$$

$$(PST-(\Phi_1 PST+PSD_1 T)B+(\Phi_1 PSD_1 T)B^2)\mathbf{z}_t=(I-\Theta_1 B)\mathbf{b}_t$$

and so, premultiplying by the "transformation" matrix PST, we obtain

$$(I-((PST)^{-1}\Phi_1 PST+T^{-1}D_1 T)B+((PST)^{-1}\Phi_1 PSD_1 T)B^2)\mathbf{z}_t=$$

$$(I-((PST)^{-1}\Theta_1 PST)B)\mathbf{a}_t$$

(where $\mathbf{a}_t=(PST)^{-1}\mathbf{b}_t$). Thus we are left with a VARMA(2,1) model for our original data, derived from our estimated model ([15]). We can calculate the likelihood of this final model from that of our estimated one, as discussed in §3.5.1(a). As in previous sections we only have the "independently adjustable parameters" that were estimated in the original model – at most 20, so that we can calculate $AIC_{[3.9]}$ for model [15] on the original data scale (see Table 6.10).

From above, we also have model [14] for the normalised but undifferenced data, in terms of $T^{[14]}S^{[14]}\mathbf{z}_t$ (S is a normalising matrix and T the transformation matrix from T&T). We can estimate this model and transform it back in terms of the original data $\mathbf{z}_t$. The fitting results for this are also given in Table 6.10.

Table 6.10 Model fitting information for US Hog data.

Model	LL	Npar	$AIC_{[3.9]}$
[14]	-2082.7	21	4207.3
[15]	-2089.9	15	4209.8

6.4.2 Model comparison

Table 6.10 shows that the two models both fit the data equally well – the residuals from both are satisfactory with no large outliers or correlation and there does not appear to be much to choose between the models.

We have re-estimated these two models on the first 71 observations in order to forecast the last 10 years' data. The forecasts produced by these are shown in Figure 6.8 for each component series where it appears that model [15] may produce slightly better forecasts than model [14], although they can both be quite poor on certain components. Also shown in this figure are the squared forecasting errors for component 1 produced by both models – again model [15] appears to give more accurate forecasts. In Figure 6.9 we plot the forecasts for the third component from each model together with their estimated ±2 standard error limits

Figure 6.8 Forecasts for US Hog data

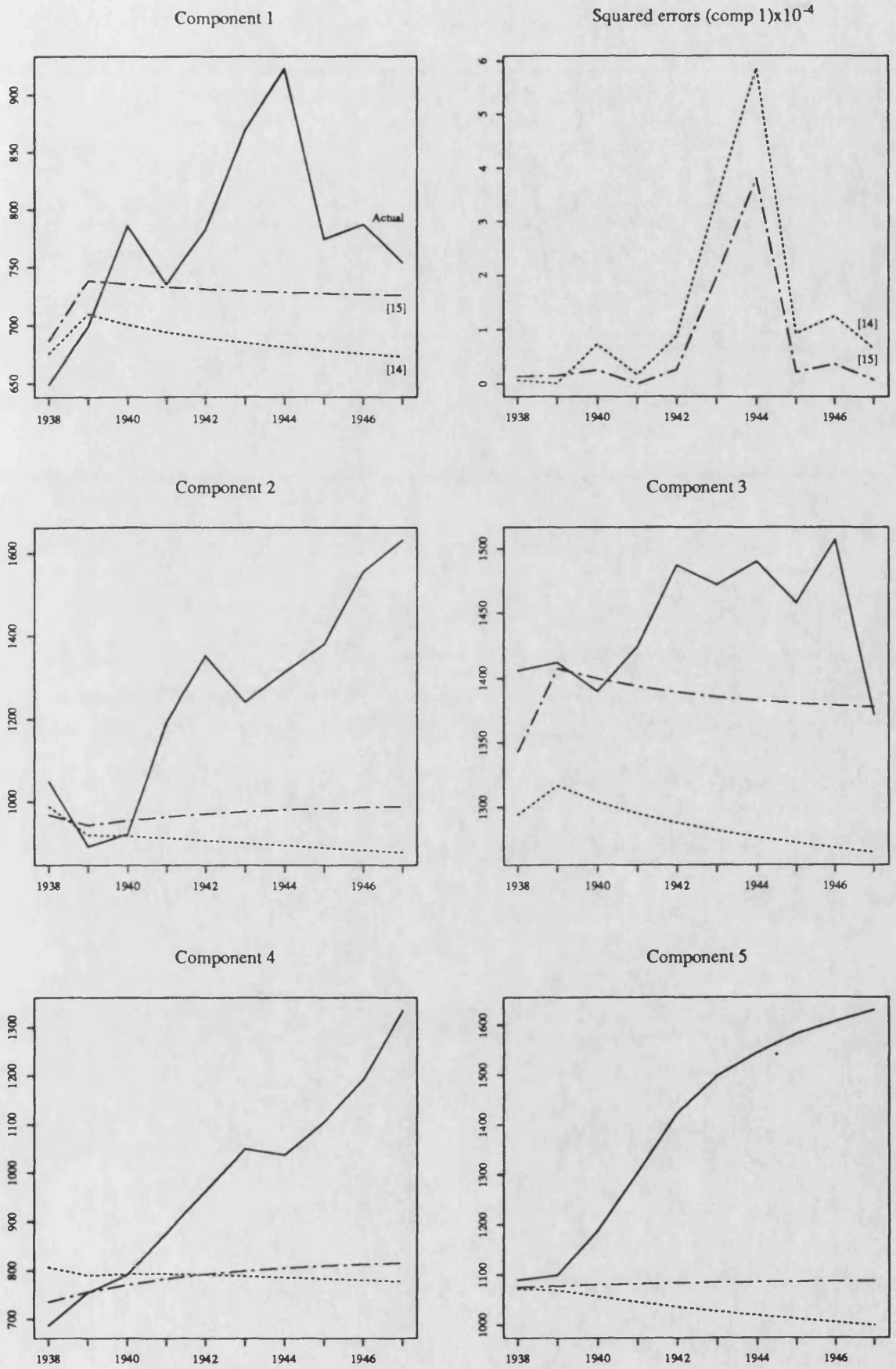
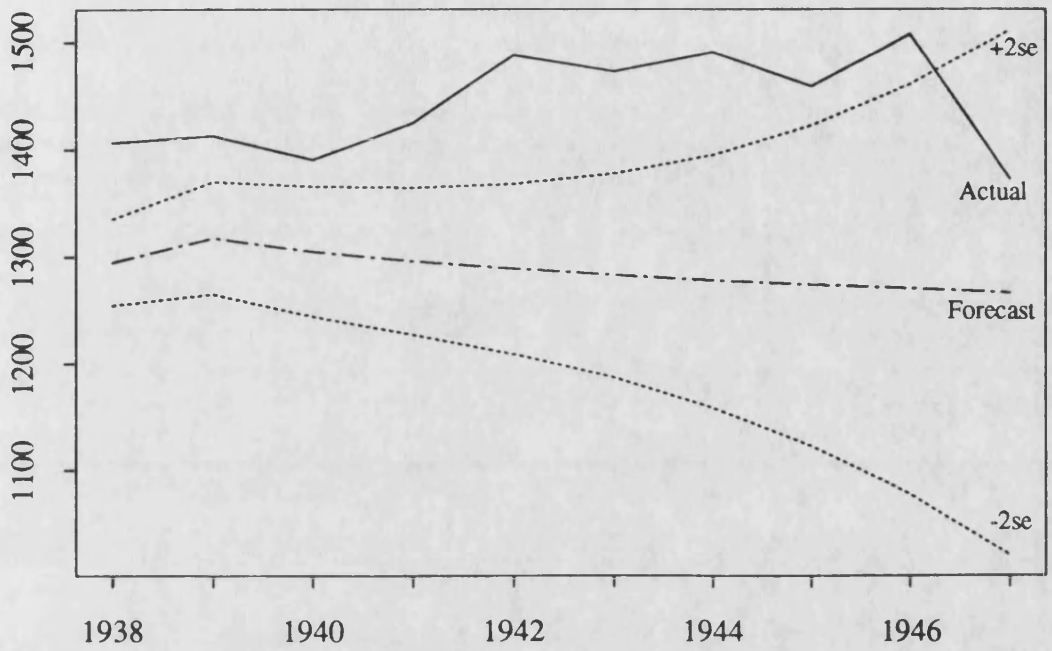
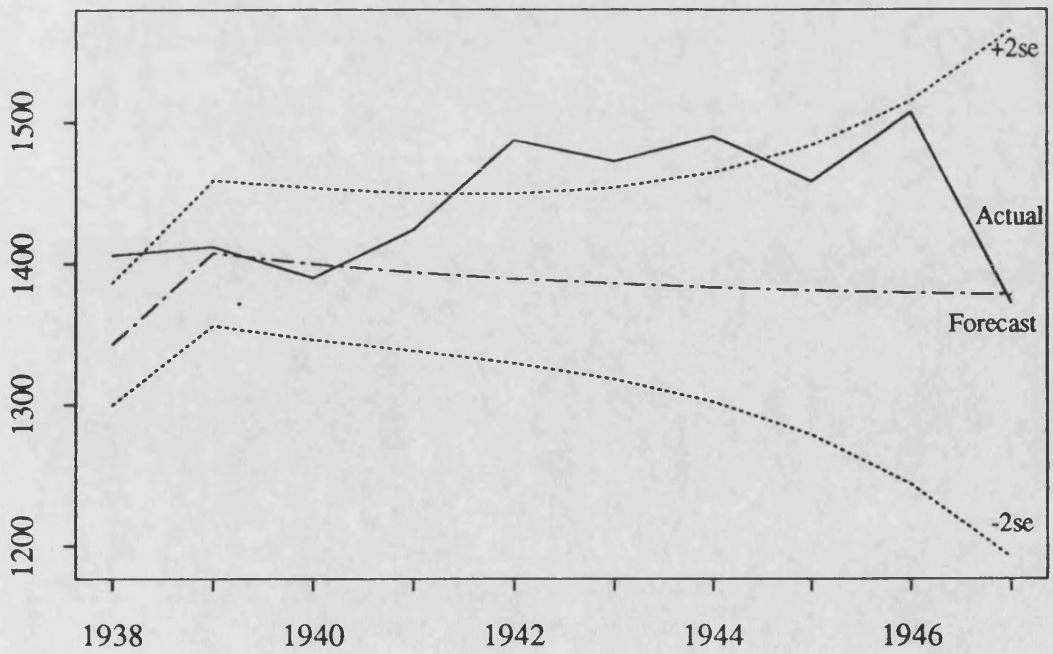


Figure 6.9

Forecasts of component 3 (Model [14])



Forecasts of component 3 ([model 15])



(calculated from the models). These plots are typical of the other components and we can see that model [15] has wider error intervals than model [14] (since it includes a n.s. term) and they capture more of the observed values. The forecasts are generally poor for this dataset, mostly because of the nature of certain of the components which change through time and the effects of the interventions (§A.4).

The two models built for the US Hog data (both using the model specification procedure of T&T) fit the data equally well, while the one built after correcting for the CI found in the data (model [15]) appears to produce somewhat superior out–of–sample forecasts. We have not used the "standard" model identification tools for comparison on this dataset since they become very complex to use for large k (see §6.3) and the data are of quite poor quality (see §A.4) – an unrestricted VARMA model is unlikely to provide an adequate representation of these relationships and would be difficult to estimate.

6.5 Conclusions

Model specification procedures such as that of T&T seem to be able to produce better VARMA models than we would usually build by simply using the cut–offs in correlation matrices to choose model orders. This is particularly true as the number of series increases (e.g. §6.3) and the cut–offs become more difficult to spot (see e.g. Table 6.8a). For more complex/higher dimension models good, parsimonious specification becomes essential to ensure accurate and fast parameter estimation. For smaller models the search path of T89b (§5.5.5) may not be more efficient than that of T&T. Also the treatment of exchangeable models is more natural within the formulation of T&T and although we do not have a way of choosing between models without estimating each of them, it is seen in §6.1.7 that the substantially different characteristics exhibited makes it essential to consider the alternative models.

As seen in §6.2, improved out–of–sample forecasting performance can be obtained from the more parsimonious models, although the extent to which this depends upon the degree of differencing of the series needs further investigation. §6.4 highlighted the need to consider carefully the (power) transformation of the data to use. Also in this section we examined the models we can obtain by analysing the raw data and data which have been made stationary by taking a CI transformation as discussed in Chapter 4. Both models were specified by model specification procedures (since the number of the variables was large) and it appeared that the model built with the stationary data had slightly superior out–of–sample forecasting performance to that for the raw data, whilst they both fitted the data equally well. The effect of this differencing/CI analysis on model behaviour clearly needs to be investigated further, but it appears to be a useful procedure to

apply to avoid difficulties with unit roots in the models to be estimated and, as discussed in §4.7.1 it imposes necessary constraints on the forecasts from such models..

Chapter 7. Concluding remarks

7.1 Summary of findings

This thesis investigates the modelling of multivariate time series. Chapter 1 discusses the nature and goals of such modelling while Chapter 2 presents the class of VARMA models which is sufficiently flexible to describe many different dynamic relationships. Although Chapter 2 discusses existing techniques for building such models, many outstanding problems were identified. They include choosing a suitable model and restricting the parameterisation to allow accurate estimation of the coefficients (both of which can be difficult with large datasets – see e.g. §6.3). Chapter 3 discussed some of these and other problems of VARMA modelling.

Co–integration (=CI) can also be a feature of interacting non–stationary variables and in §4.2 the effects of this on unrestricted VARMA models were considered. The techniques of Engle and Granger (1987) (=E&G), discussed in Chapter 4, can be used to explore CI and §4.4.2 presents a new strategy for using these techniques to construct suitable transformations and differencing operators which can be applied to correct for CI within VARMA models. The case study in §4.5 demonstrated this strategy and the CI relationships which were found may themselves prove interesting. Extensions of the methods to higher degree differencing are straightforward. The forecasts from CI models remain linked in an appropriate way, which should prove crucial in improving the forecasting performance of such models – it was seen in §6.4 that the CI model [15] was able to produce better forecasts than the unrestricted model [14].

Some related approaches to VARMA model specification were discussed in Chapter 5. These aim to identify a suitable model and specify a parsimonious parameterisation to improve estimation and modelling. Some detailed aspects of the procedures were discussed including the close relationships between them and their relative performance at finding various order SCMs. It was seen that each of the methods looks for patterns in the same constructed covariance matrices, but one method, denoted by T89a, cannot find mixed models and so must usually be discounted. T89b extends the method of T89a and can find mixed models. It is also substantially quicker than the method of Tiao and Tsay (1989) (=T&T) for larger order models since it does not introduce redundancy at every stage (although T89b may take longer to find collections of smaller models). However, the treatment of exchangeable or alternative models is more natural within T&T and the case studies in §6.1 and §6.2 demonstrate that the models may have substantially different characteristics, so that it is important to entertain all the

alternative models. This consideration suggests that T&T provides a fuller exploration of the SCM structure than T89b and so, although T&T is necessarily more complex, it may often be preferable since it can find alternative models. There does not appear to be an immediate way of choosing between exchangeable models – perhaps the analysis must be performed with each possible model and comparisons made in order to choose the one most suitable for the application.

7.2 Strategy for model-building

The primary aim of this thesis has been to develop a sensible general strategy for building VARMA models. This section summarises the recommended strategy.

Initial examination of data

Much of the advice from the univariate case, such as plotting the data, seeking relevant external information and applying exploratory techniques, also applies to the multivariate case (see e.g. Jenkins (1979) or Mills (1990, Part 1)). A univariate analysis of each (or selected) component series can often be a useful first stage. The collection of univariate models also provides a useful "benchmark" with which to compare multivariate models.

Non-stationarity

In the multivariate case it is recommended to first test n.s. series for CI (§4.4.2) in order to specify the "minimal" differencing operator (§4.2.3), which will avoid estimation problems. It is preferable to work with stationary data, but without inducing unit roots in the MA polynomial, as discussed in §4.2. The CI relationships themselves may prove interesting and should certainly be applied to constrain any forecasts as discussed earlier (also §4.7.1).

VARMA model specification

Datasets comprising two or perhaps three series, with simple structures, can often be modelled well by using the "standard" techniques of Chapter 2. However, higher dimension or more complex structures are difficult to specify and (particularly if redundancy is likely to be present) it becomes essential to ensure a parsimonious parameterisation by using a model specification procedure such as T&T or T89b. These procedures are an extension of the "model identification" stage of §1.2.2 and they attempt to specify **minimal components** of the multivariate structure and provide initial estimates of the parameters in order to give a more accurate estimation. It has been seen that T&T may be preferable to T89b since it can suggest alternative structures, any of which could be chosen depending on the application. These procedures are complicated and can be

difficult to implement. However, with a suitable implementation, the time spent using the procedures can be relatively small. For more than about 8 component series, VARMA model building becomes too complex and time–consuming, so that some form of dimension reduction or analysis of suitable subsets of the components becomes necessary.

Model estimation and checking

Having chosen suitable VARMA model(s) for a dataset and estimated the parameters by maximum likelihood (iteratively deleting insignificant ones), the diagnostic checks may suggest inadequacies which demand different models or maybe even alternative representations. Perhaps a non–linear or time–varying parameter model might be needed to better describe the relationships.

Other analyses

The coefficients of the fitted VARMA model may also suggest other possible analyses/approaches. For example, the form of the model (or perhaps its SCMs) may indicate that dimension reduction is possible, or that a transfer function representation would be sufficient. Application of these more specialised analyses is likely to give superior results to the more general VARMA model in situations for which they are suitable.

7.3 Further Research

Many avenues for further research suggest themselves, particularly those concerned with comparative studies of various techniques.

The alternative approaches to CI, briefly discussed in §4.7.3, need to be studied, including other test procedures, in view of the criticisms of the testing of E&G. With regard to model specification, the preliminary comparisons from Chapter 6 suggest that T&T can handle exchangeability more easily than T89b, although it might be better to incorporate the more efficient search path of T89b into T&T in some way. Other approaches to model specification (mentioned in §5.6) require assessment and comparison with e.g. T&T, in particular to determine if T&T can be used to meet the objectives of these other approaches (such as dimension reduction). Further work is needed on modelling seasonality. It is inevitable that seasonal modelling will be much more complex than for the non–seasonal case.

Some additional technical topics for further investigation include (1) an extension of the model metric idea of §3.3 for model comparison, (2) the choice of criteria for comparing multivariate forecasts and (3) the effect of the value of the variable "h" on the method of T&T (see §5.2.7).

An important practical issue concerns the extent to which multivariate models can be applied to meet the modelling objectives. Such objectives include understanding and describing the interrelationships between the observed variables, the within–sample explanatory capacity and the (out–of–sample) forecasting performance.

Appendix: Data

In this Appendix we present some datasets which have been used in the text for illustration. Background and previous analyses are discussed in order to understand the nature and any shortcomings of the data.

A.1 Lydia Pinkham data

Palda (1964) gives a background to this dataset which we include for interest and to highlight some historical points. Lydia Pinkham's Vegetable Compound was a herbal extract brewed by the Lydia Pinkham Company in the US and first sold in 1873. A court case made available high quality data on the company's sales and expenditure on advertising: 54 annual observations from 1907 to 1960; 78 monthly observations from January 1954 to June 1960 (Palda (1964)). These data are plotted in Figures A.1a and b and tabulated in Tables A.1a and b. Throughout its history, the company spent a large proportion of its profit on advertising the product. Almost all of this was spent on newspaper copy, there were only a few, small (deflated, presumably) price changes during the period of the data, no direct competitors to the product and a stable distribution system. For these reasons the data is remarkably well suited to analysis to measure the effect of advertising expenditure on sales. Such analyses have been carried out quite extensively and many publications report models built attempting to describe this effect (e.g. see Kyle (1978), Helmer and Johansson (1977)). The data also suggest some feedback effect (which is not unreasonable, since advertising expenditure may be determined as a proportion of available profits and hence sales) and this has also been modelled (e.g. Hanseens (1980), Bhattacharyya (1982)). Heyse and Wei (1985b) discuss some of these (essentially transfer function) models and present a VARMA analysis of the annual data. The possible existence of feedback and ignorance of the dynamic relationships make the VARMA model a flexible and sensible choice. Heyse and Wei's model suggests that this year's sales affect advertising next year and that advertising this year does not affect sales next, with the largest effect being contemporaneous. Clarke (1976) presented a comprehensive discussion of the **duration** of advertising's effect on sales and concluded that for mature, frequently purchased, low–priced products, (which the Lydia Pinkham compound certainly was) the effect lasts between 3 and 15 months with 90% of the effect within 3 to 9 months of the advertising effort. In §6.1 we apply the methods of Chapters 4 and 5 to see if their use can improve the analysis.

Possible **explained** interventions in the annual data which may have altered the way in which sales responded to advertising expenditure include:

1914: change in the solid content of the product

1925: weaker medical claims made in advertising

1941: stronger medical claims made in advertising

(these last two changes follow efforts by the US authorities to make the company back–up its, often exaggerated advertising by demonstrating the medicinal properties of the product). Note that Bhattacharyya (1982) fits 7 interventions in models for the **monthly** data in order to control some outlying residuals.

A.2 Flour price data

T&T used a previously unpublished dataset (as well as the US Hog data – §A.4) to illustrate the application of their model specification procedure (§5.2). These data consist of an index of the price of flour, measured monthly at three US cities – Buffalo, Minneapolis and Kansas City – from August 1972 to November 1980 (100 observations). They are plotted in Figure A.2 and tabulated in Table A.2. Throughout their analysis, T&T used the logged data (we display the raw form) and in the analysis of §6.2 we will also work with this transform to allow comparisons to be made. In the range of the data (100 – 200) the log function is nearly linear, so that the time plots of the logged data preserve the features of Figure A.2. The data have not apparently been seasonally adjusted and there is no evidence of seasonality, simply of strong non–stationarity, with the series each resembling univariate random walks (as many price series do). The three series all follow one another closely, suggesting strong interactions, although the largest effect is seen to be simultaneous (perhaps, not surprisingly, since a price difference between cities could easily be exploited within a month, given the proximity of the cities). The analysis is carried out with a view to describing and explaining the dynamic relationships between the series, rather than specifically forecasting future price levels. We build and compare various models for this data in §6.2.

A.3 UK Pig production data

Andrews and Herzberg (1985) tabulate some data on UK Pig production. It consists of quarterly observations on the five variables detailed below.

z_{1t} Gilts: a measure of the intake of pigs into the breeding herd.

z_{2t} Profit: an index of profit made on the sale of pigs.

z_{3t} Slaur: ratio of sow and boar slaughters to the total breeding herd size (a measure of the removal of pigs from the breeding herd).

z_{4t} Clean: a measure of the number of clean pigs reared for meat (not breeding).

z_{5t} Herd: a measure of the actual herd size.

These data are tabulated in Table A.3 and plotted in Figure A.3. As with the US Hog data in §A.4, the sampling schemes used to measure the variables change slightly over the period of interest. In §6.3 we seasonally adjust the series before applying the model specification procedures to them.

A.4 US Hog data

Quenouille (1968) analysed some data on the US Hog industry which consisted of five annual variables measured from 1867 to 1948 (82 observations). The variable definitions are given below.

Description	Name	Basic variable	Variable used
Number of hogs	Hogs	h_n = number of hogs recorded on 1st January in the Census of Agriculture	$z_{1t}=1000(\log_{10}h_n-7)$
Price of hogs	Hog$	$h_\$$ = $value/head	$z_{2t}=1000\log_{10}h_\$$
Price of corn	Corn$	$c_{\$,1}$ = $price/bushel for all purposes on 1st December†	$z_{3t}=1000(\log_{10}c_{\$,1}+1)$
		$c_{\$,2}$ = average seasonal price†	$z_{3t}=1000(\log_{10}c_{\$,2}+1)$
Supply of corn	Corn	c_n = number of bushels produced for all purposes	$z_{4t}=1000(\log_{10}c_n-8)$
Farm wage rate	Farm$	$w_\$$ = a measure of the rate of farm wages, normalised to 100 in 1910–1914‡	$z_{5t}=1000(\log_{10}w_\$-1)$

† Neither of these measures were available during the whole period, so the first was adjusted to be comparable with the second during the years for which they were both available, the variables used are then $c_{\$,1}+51.1$, (1867–1907) and $c_{\$,2}$, (1908–1948).

‡ Many of the observations have been produced by linear interpolation – see the plot in Figure A.4e.

These data are tabulated in Table A.4 and plotted in Figure A.4. Quenouille analysed $(z_{1t},z_{2t},z_{3t},z_{4t},z_{5t})^T$, whilst Box and Tiao (1977) suggested that using the data $(z_{1t},z_{2,t+1},z_{3t},z_{4t},z_{5,t+1})^T$ could improve the fit of their models. T&T claim to analyse $\mathbf{z}_t=(z_{1t},z_{2,t+1},z_{4t},z_{3t},z_{5,t+1})^T$, although the models fitted suggest (and Tsay (personal communication, 1989) confirms) that T&T actually analysed $\mathbf{y}_t=\ln(\mathbf{z}_t)$ in order to reduce the scale of the series. This is considered in §6.4. In our analyses we use $\mathbf{z}_t$ as above.

There are clearly interventions in the data (e.g. the Great Depression of the 1930s) which together with the notes † and ‡ above makes them a difficult set of series to analyse. They are perhaps best suited to smaller models exploring the effects between for example the pair of price series. In §4.5 we analyse the data to find any CI relationships between the variables (which are n.s. – see e.g. Figure A.4).

A.5 Other data

There are many other published multivariate time series datasets that we could use to compare the various methods, particularly in the further studies suggested in §7.3, but the four chosen datasets presented above are sufficient for our purposes in this thesis.

Table A.1a Annual Lydia Pinkham data (in $1000).

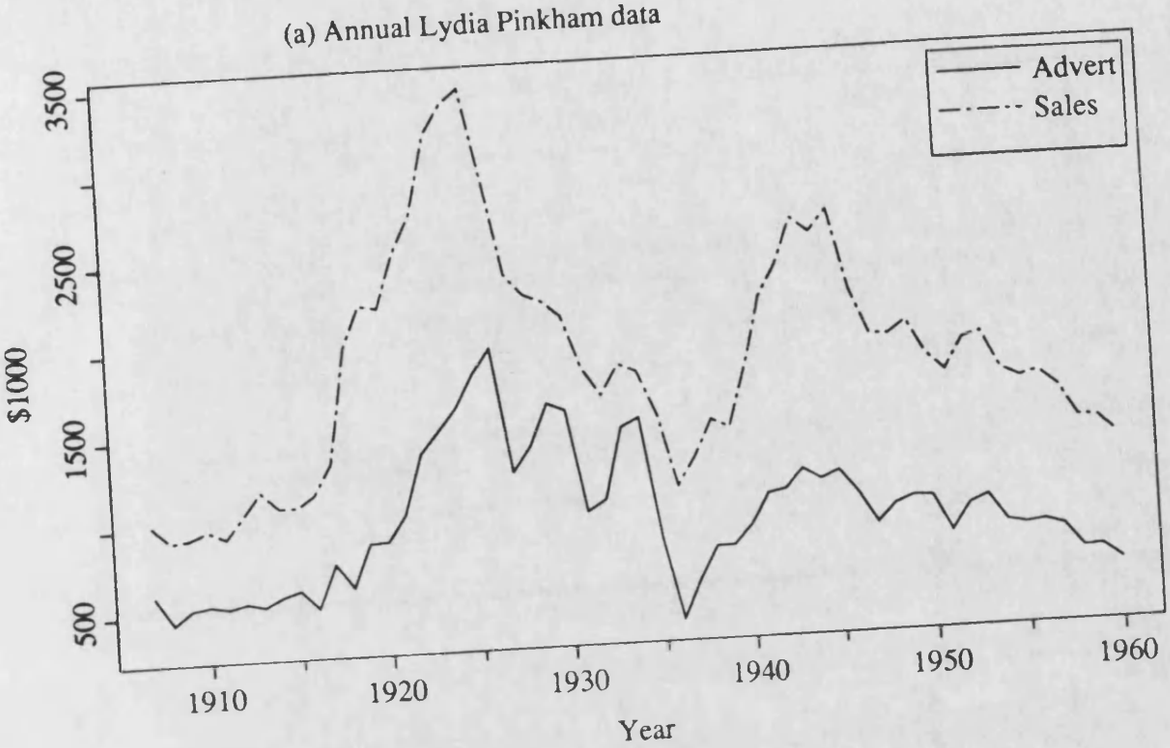
Year	Adv	Sal	Year	Adv	Sal
1907	608	1016	1934	1504	1770
1908	451	921	1935	807	1518
1909	529	934	1936	339	1103
1910	543	976	1937	562	1266
1911	525	930	1938	745	1473
1912	549	1052	1939	749	1423
1913	525	1184	1940	862	1767
1914	578	1089	1941	1034	2161
1915	609	1087	1942	1054	2336
1916	504	1154	1943	1164	2602
1917	752	1330	1944	1102	2518
1918	613	1980	1945	1145	2637
1919	862	2223	1946	1012	2177
1920	866	2203	1947	836	1920
1921	1016	2514	1948	941	1910
1922	1360	2726	1949	981	1984
1923	1482	3185	1950	974	1787
1924	1608	3351	1951	766	1689
1925	1800	3438	1952	920	1866
1926	1941	2917	1953	964	1896
1927	1229	2359	1954	811	1684
1928	1373	2240	1955	789	1633
1929	1611	2196	1956	802	1657
1930	1568	2111	1957	770	1569
1931	983	1806	1958	639	1390
1932	1046	1644	1959	644	1387
1933	1453	1814	1960	564	1289

Table A.1b Monthly Lydia Pinkham data (in $100).

Year	Month	Sal	Adv	Year	Month	Sal	Adv
1954	1	1295	1280		4	1534	1012
	2	1318	1350		5	1332	745
	3	1728	982		6	1200	78
	4	1539	919		7	1314	66
	5	1324	87		8	1180	94
	6	1264	39		9	1264	774
	7	1169	72		10	1318	971
	8	1479	467		11	1018	536
	9	1631	1170		12	1438	150
	10	1546	917	1958	1	772	580
	11	1459	701		2	902	1121
	12	1087	128		3	1265	974
1955	1	1171	1014		4	1229	1002
	2	1406	1274		5	1318	138
	3	1619	1388		6	1195	72
	4	1508	1071		7	1105	59
	5	1521	537		8	1095	270
	6	1341	123		9	1298	986
	7	1247	60		10	1482	673
	8	1262	351		11	1163	304
	9	1419	1061		12	1072	209
	10	1558	791	1959	1	1052	838
	11	1222	138		2	1102	994
	12	1053	77		3	1355	1020
1956	1	1242	1000		4	1323	865
	2	1361	1182		5	1296	819
	3	1660	1225		6	1127	83
	4	1717	936		7	1170	56
	5	1371	625		8	1059	224
	6	1293	60		9	1116	881
	7	1285	61		10	1214	436
	8	1210	169		11	966	160
	9	1142	946		12	1089	68
	10	1586	1306	1960	1	814	749
	11	1441	426		2	1087	857
	12	1262	88		3	1180	898
1957	1	1267	1104		4	1167	705
	2	1278	1093		5	1210	489
	3	1544	1080		6	1092	59

Figure A.1

(a) Annual Lydia Pinkham data



(b) Monthly Lydia Pinkham data

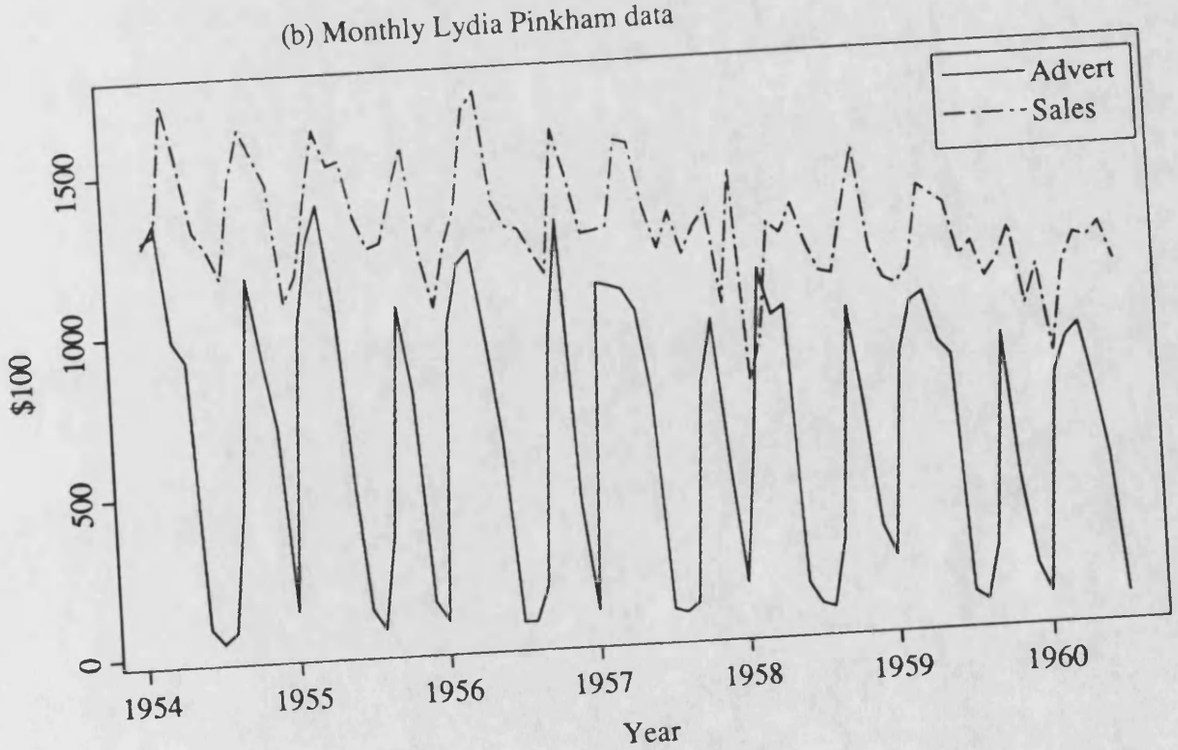
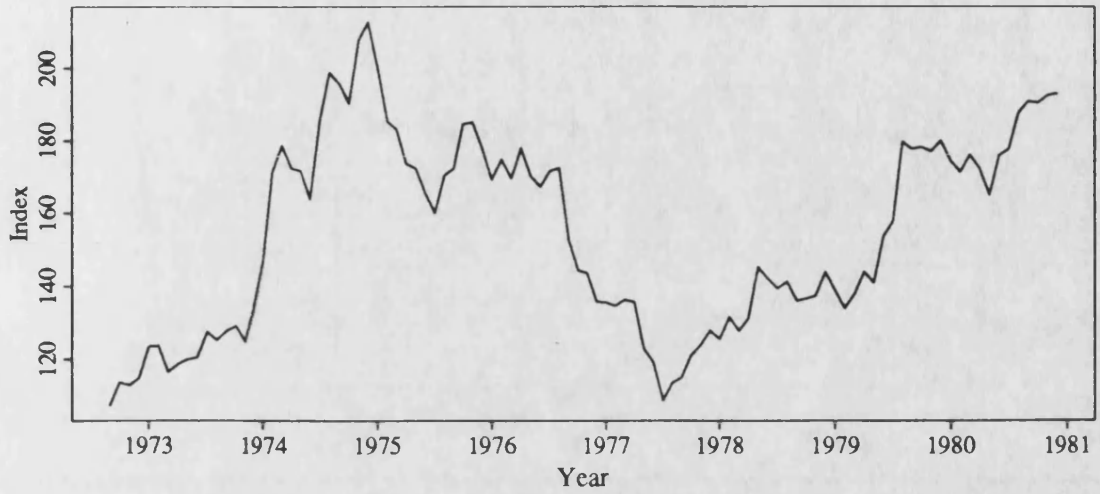


Table A.2 Flour Price data.

Year	Month	Buff	Minn	Kans	Year	Month	Buff	Minn	Kans
1972	8	107.1	106.5	110.9		10	143.5	137.3	134.2
	9	113.5	112.4	114.6		11	135.6	129.7	126.1
	10	112.7	111.8	115.5		12	135.4	128.4	124.2
	11	114.7	113.3	117.0	1977	1	134.5	126.9	122.7
	12	123.4	124.5	135.0		2	136.1	128.8	123.5
1973	1	123.6	124.3	132.8		3	135.6	126.5	118.3
	2	116.3	116.5	122.6		4	122.8	116.6	112.3
	3	118.5	118.6	123.8		5	119.0	113.4	105.7
	4	119.8	119.6	128.9		6	108.5	102.8	97.7
	5	120.3	119.4	126.7		7	113.3	107.7	105.8
	6	127.4	128.6	139.3		8	114.8	109.4	106.9
	7	125.1	126.3	135.7		9	120.9	114.9	110.0
	8	127.6	126.8	135.6		10	123.7	117.5	114.3
	9	129.0	125.7	146.0		11	127.8	120.0	118.8
	10	124.6	120.8	140.7		12	125.4	117.6	117.2
	11	134.1	127.9	147.0	1978	1	131.5	124.0	126.1
	12	146.5	147.6	163.9		2	127.7	119.7	120.5
1974	1	171.2	169.8	194.3		3	131.2	125.0	125.6
	2	178.6	177.6	200.8		4	145.2	141.1	132.0
	3	172.2	172.5	193.4		5	141.9	137.0	134.6
	4	171.5	170.1	190.3		6	139.3	132.3	130.3
	5	163.6	171.3	188.0		7	141.1	134.8	137.0
	6	185.6	189.9	196.1		8	135.9	129.7	136.6
	7	198.8	206.9	215.0		9	136.5	128.7	137.0
	8	195.7	197.4	201.6		10	137.2	129.9	138.4
	9	190.3	195.0	203.4		11	143.8	137.2	142.9
	10	207.9	214.2	222.1		12	138.7	132.8	140.4
	11	212.8	219.2	228.7	1979	1	133.9	127.5	136.0
	12	199.9	205.6	216.1		2	137.7	131.2	140.1
1975	1	185.3	193.4	200.2		3	143.8	137.1	148.2
	2	183.0	185.1	189.6		4	140.8	135.5	146.4
	3	173.5	174.0	173.3		5	153.4	147.1	158.5
	4	172.2	173.2	169.7		6	157.5	151.6	163.5
	5	165.3	164.5	161.0		7	179.5	173.7	187.1
	6	159.9	158.9	151.7		8	177.5	171.6	181.7
	7	170.3	169.7	167.1		9	178.0	170.8	181.5
	8	172.2	174.4	174.4		10	176.8	172.4	181.9
	9	184.5	186.2	189.7		11	179.8	174.9	190.9
	10	185.0	184.7	187.4		12	174.2	168.1	186.9
	11	177.7	176.4	178.4	1980	1	171.1	164.7	180.1
	12	169.1	167.6	165.8		2	175.9	170.0	184.8
1976	1	174.7	170.9	164.9		3	172.2	164.9	174.8
	2	169.4	168.3	171.8		4	164.7	157.9	169.0
	3	177.8	176.4	175.4		5	175.7	169.2	178.4
	4	170.1	168.6	165.9		6	177.4	168.6	175.3
	5	167.1	164.6	157.3		7	187.5	179.8	178.2
	6	171.4	170.1	161.4		8	190.7	179.0	182.0
	7	172.3	169.4	159.2		9	190.4	179.2	188.6
	8	152.6	149.6	142.8		10	192.4	181.4	190.8
	9	144.1	139.5	138.5		11	192.9	181.8	192.2

Figure A.2 Flour price data

(a) Buffalo



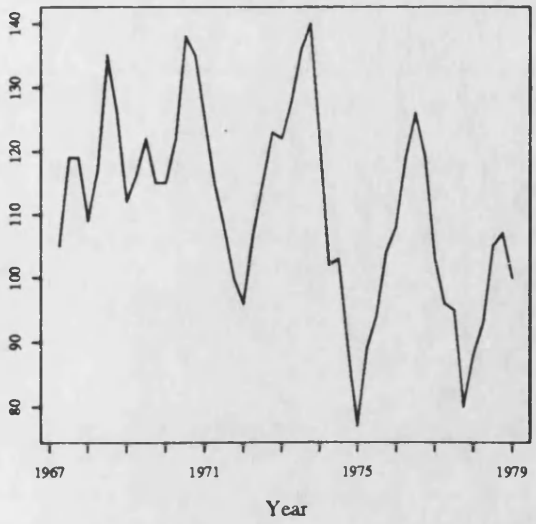
(b) Minneapolis

(c) Kansas

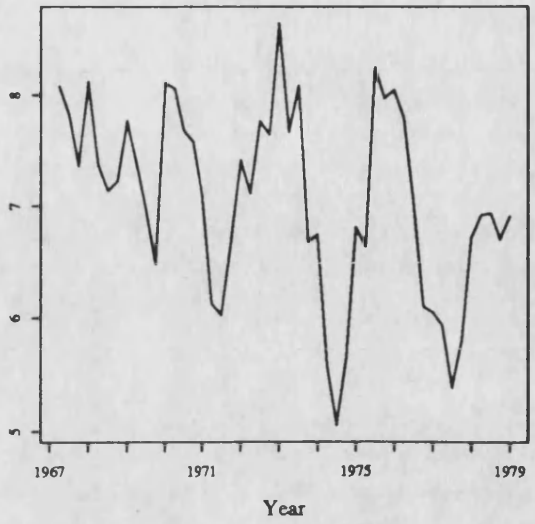
Table A.3 UK Pig production data.

Year	¼	Gilts	Profit	Slaur	Clean	Herd
1967	1	105	8.075	10.80	2645	703
	2	119	7.819	9.16	2540	722
	3	119	7.366	9.38	2565	738
	4	109	8.113	10.39	2776	747
1968	1	117	7.380	9.44	2725	755
	2	135	7.134	8.69	2623	780
	3	126	7.222	9.60	2722	806
	4	112	7.768	11.28	3004	807
1969	1	116	7.386	11.20	2952	805
	2	122	6.965	9.94	2968	801
	3	115	6.478	11.21	2961	821
	4	115	8.105	11.69	3243	809
1970	1	122	8.060	9.67	3027	797
	2	138	7.684	7.87	2902	831
	3	135	7.580	8.15	3057	867
	4	125	7.093	8.83	3331	862
1971	1	115	6.129	10.51	3266	871
	2	108	6.026	9.03	3290	864
	3	100	6.679	9.93	3223	854
	4	96	7.414	10.27	3501	846
1972	1	107	7.112	9.56	3402	854
	2	115	7.762	8.74	3278	851
	3	123	7.645	10.32	3258	876
	4	122	8.639	10.31	3400	876
1973	1	128	7.667	9.97	3303	888
	2	136	8.080	8.99	3228	903
	3	140	6.678	12.22	3269	922
	4	122	6.739	12.90	3396	902
1974	1	102	5.569	14.00	3396	820
	2	103	5.049	12.77	3386	819
	3	89	5.642	12.61	3385	797
	4	77	6.808	12.16	3262	751
1975	1	89	6.636	11.16	3113	743
	2	94	8.241	8.90	2851	744
	3	104	7.968	10.24	2752	747
	4	108	8.044	10.29	2919	764
1976	1	119	7.791	10.03	2842	759
	2	126	7.024	9.05	2834	807
	3	119	6.102	12.00	2957	798
	4	103	6.053	12.70	3305	811
1977	1	96	5.941	12.95	3256	752
	2	95	5.386	11.69	3151	761
	3	80	5.811	12.65	3141	719
	4	88	6.716	12.21	3266	741
1978	1	93	6.923	10.68	3061	745
	2	105	6.939	10.47	3018	764
	3	107	6.705	11.05	3085	764
	4	100	6.914	10.31	3242	786

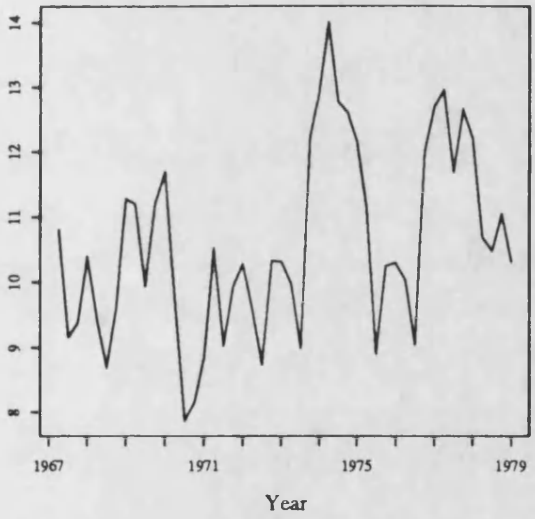
(a) Gilts



(b) Profit



(c) Slaur



(d) Clean

(e) Herd

Figure A.3

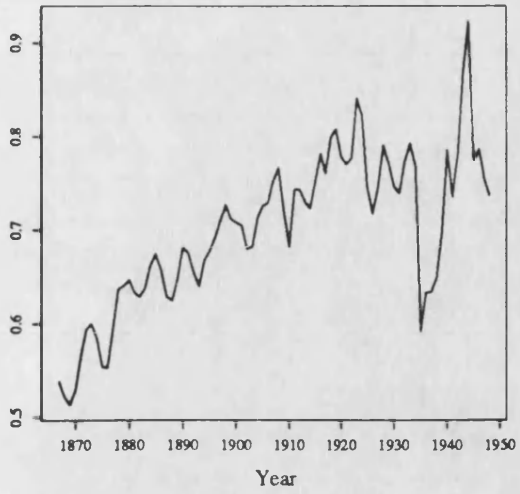
UK Pig production data

Table A.4 Quenouille's US hog data.

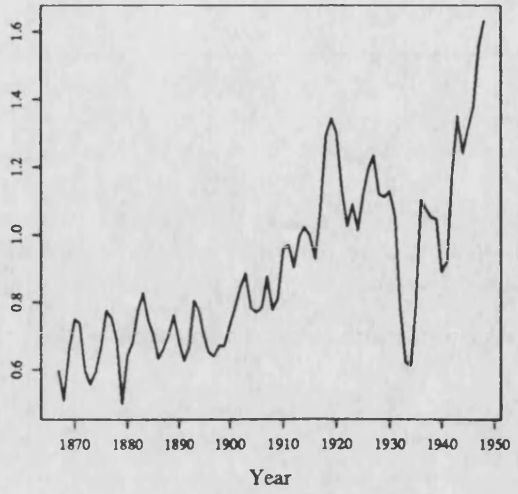
Year	Hogs	Hog$	Corn$	Corn	Farm$
1867	538	597	944	900	722
1868	522	509	841	964	719
1869	513	663	911	893	716
1870	529	751	768	1051	724
1871	565	739	718	1057	732
1872	594	598	634	1107	740
1873	600	556	735	1003	748
1874	584	594	858	1025	756
1875	554	667	673	1161	748
1876	553	776	609	1170	740
1877	595	754	604	1181	732
1878	637	689	457	1194	744
1879	641	498	612	1244	756
1880	647	643	642	1232	778
1881	634	681	849	1095	799
1882	629	778	733	1244	799
1883	638	829	672	1218	799
1884	662	751	594	1289	799
1885	675	704	559	1313	801
1886	658	633	604	1251	803
1887	629	663	678	1205	806
1888	625	709	571	1352	806
1889	648	763	490	1361	806
1890	682	681	747	1218	810
1891	676	627	651	1368	813
1892	655	667	645	1278	810
1893	640	804	609	1279	806
1894	668	782	705	1208	771
1895	678	707	453	1404	771
1896	692	653	382	1427	780
1897	710	639	466	1359	789
1898	727	672	506	1371	799
1899	712	669	525	1423	820
1900	708	729	595	1425	834
1901	705	784	829	1234	848
1902	680	842	654	1443	863
1903	682	886	673	1401	884
1904	713	784	691	1429	906
1905	726	770	660	1470	928
1906	729	783	643	1482	949
1907	752	877	754	1417	960
1908	766	777	813	1409	971
1909	720	810	790	1417	982
1910	682	957	712	1455	987
1911	743	970	831	1394	991
1912	743	903	742	1469	1004
1913	730	995	847	1357	1013
1914	723	1022	850	1402	1004
1915	753	998	830	1452	1013
1916	782	928	1056	1385	1053
1917	760	1073	1163	1464	1149

Year	Hogs	Hog$	Corn$	Corn	Farm$
1918	799	1294	1182	1388	1248
1919	808	1346	1180	1428	1316
1920	779	1301	805	1487	1384
1921	770	1134	714	1467	1190
1922	777	1024	865	1432	1179
1923	841	1090	911	1459	1228
1924	823	1013	1027	1347	1238
1925	746	1119	846	1447	1246
1926	717	1195	869	1406	1253
1927	744	1235	928	1418	1253
1928	791	1120	924	1426	1253
1929	771	1112	903	1401	1255
1930	746	1129	777	1318	1223
1931	739	1055	507	1411	1114
1932	773	787	500	1467	982
1933	793	624	716	1380	929
1934	768	612	911	1161	978
1935	592	800	816	1362	1013
1936	633	1104	1019	1178	1045
1937	634	1075	714	1422	1100
1938	649	1052	687	1406	1097
1939	699	1048	754	1412	1090
1940	786	891	791	1390	1100
1941	735	921	876	1424	1188
1942	782	1193	962	1487	1303
1943	869	1352	1050	1472	1422
1944	923	1243	1037	1490	1498
1945	774	1314	1104	1458	1544
1946	787	1380	1193	1507	1582
1947	754	1556	1334	1372	1607
1948	737	1632	1114	1557	1629

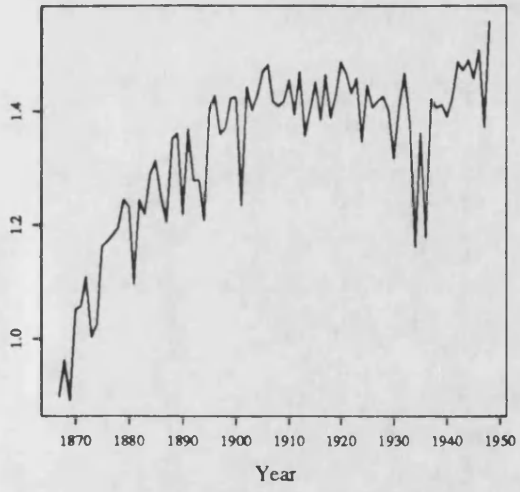
(a) Hogs/1000



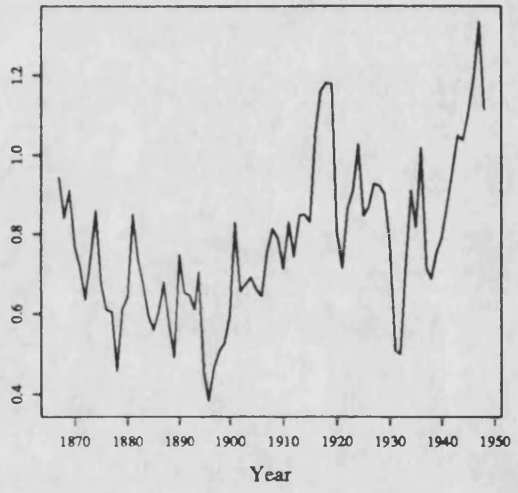
(b) Hog$/1000



(c) Corn/1000



(d) Corn$/1000



(e) Farm$/1000

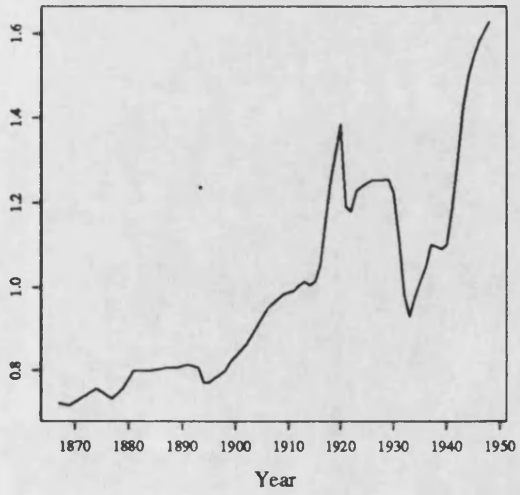


Figure A.4

Quenouille's US Hog data

References

Ahn, S.K. and Reinsel, G.C. (1988) Nested reduced–rank autoregressive models for multiple time series. *Journal of the American Statistical Association*, **83**, 849–856.

Akaike, H. (1976) Canonical correlation analysis of time series and the use of an information criterion. in *Systems Identification: Advances and Case Studies*. (eds. R.K. Mehra and D.G. Lainiotis) pp27–96. New York: Academic Press.

Andrews, D.F. and Herzberg, A.M. (1985) *Data*. New York: Springer–Verlag.

Ansley, C.F. and Kohn, R. (1983) Exact likelihood of a vector autoregressive moving average process with missing or aggregated data. *Biometrika*, **70**, 275–278.

Ashley, R. (1983) On the usefulness of macroeconomic forecasts as inputs to forecasting models. *Journal of Forecasting*, **2**, 211–223.

Bhattacharyya, M.N. (1982) Lydia Pinkham data remodelled. *Journal of Time Series Analysis*, **3**, 81–102.

Boero, G. (1990) Comparing *ex–ante* forecasts from a SEM and VAR model: an application to the Italian economy. *Journal of Forecasting*, **9**, 13–24.

Bossaerts, P. (1988) Common nonstationary components of asset prices. *Journal of Economic Dynamics and Control*, **12**, 347–364.

Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, **26**, 211–252.

Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. 2nd edn. San Francisco: Holden–Day.

Box, G.E.P. and Tiao, G.C. (1975) Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, **70**, 70–79.

Box, G.E.P. and Tiao, G.C. (1977) A canonical analysis of multiple time series. *Biometrika*, **64**, 355–365.

Brockwell, P.J. and Davis, R.A. (1987) *Time Series: Theory and Methods*. New York: Springer–Verlag.

Bruce, A.G. and Martin, R.D. (1989) Leave–*k*–out diagnostics for time series (with Discussion). *Journal of the Royal Statistical Society*, Series B, **51**, 363–424.

Buuren, S. van (1990) *Optimal Scaling of Time Series*. Leiden: DSWO Press.

Chang, I., Tiao, G.C. and Chen, C. (1988) Estimation of time series parameters in the presence of outliers. *Technometrics*, **30**, 193–204.

Chatfield, C (1978) The Holt–Winters forecasting procedure. *Applied Statistics*, **27**, 264–279.

Chatfield, C. (1979) Inverse autocorrelations. *Journal of the Royal Statistical Society*, Series A, **142**, 363–377.

Chatfield, C. (1988b) What is the "best" method for forecasting? *Journal of Applied Statistics*, **15**, 19–38.

Chatfield, C. (1988c) Apples, oranges and mean square error. Editorial. *International Journal of Forecasting*, **4**, 515–518.

Chatfield, C. (1989) *The Analysis of Time Series*. 4th edn. London: Chapman and Hall.

Chatfield, C. and Collins, A.J. (1980) *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Clarke, D.G. (1976) The econometric measurement of the duration of advertising's effect on sales. *Journal of Marketing Research*, **13**, 345–357.

Cleveland, W.S. and Devlin, S.J. (1982) Calendar effects in monthly time series: modelling and adjustment. *Journal of the American Statistical Association*, **77**, 520–528.

Cooper, D.M. and Wood, E.F. (1982) Identifying multivariate time series models. *Journal of Time Series Analysis*, **3**, 153–164.

Dickey, D.A. and Fuller, W.A. (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**, 427–431.

Dickey, D.A. and Fuller, W.A. (1981) The likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, **49**, 1057–1072.

Dickey, D.A., Bell, B. and Miller, R. (1986) Unit roots in time series models: tests and implications. *The American Statistician*, **40**, 12–26.

Dillon, W.R. and Goldstein, M. (1984) *Multivariate Analysis*. New York: Wiley.

Engle, R.F. and Granger, C.W.J. (1987) Co–integration and error correction: representation, estimation and testing. *Econometrica*, **55**, 251–276.

Engle, R.F., Granger, C.W.J. and Hallman, J.J. (1989) Merging short–run and long–run forecasts – an application of seasonal co–integration to monthly electricity sales forecasting. *Journal of Econometrics*, **40**, 45–62.

Engle, R.F. and Yoo, B.S. (1987) Forecasting and testing in co–integrated systems. *Journal of Econometrics*, **35**, 143–159.

Fildes, R. (1989) The evaluation of extrapolative forecasting models. Manuscript, Manchester Business School.

Fox, A.J. (1972) Outliers in time series. *Journal of the Royal Statistical Society*, Series B, **43**, 350–363.

Gooijer, J.G. de and Klein, A. (1989) Forecasting the Antwerp maritime steel traffic flow: a case study. *Journal of Forecasting*, **8**, 381–398.

Gooijer, J.G. de, Abraham, B., Gould, A. and Robinson, L. (1985) Methods for determining the order of an autoregressive–moving average process: a survey. *International Statistical Review*, **53**, 301–29.

Granger, C.W.J. (1981) Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 121–130.

Granger, C.W.J. and Newbold, P. (1986) *Forecasting Economic Time Series*. 2nd edn. Orlando: Academic Press.

Granger, C.W.J. and Weiss, A.A. (1983) Time series analysis of error–correction models. in *Studies in Econometrics, Time Series Analysis and Multivariate Statistics*. (eds. Karlin, S., Amemiya, T. and Goodman, L.A.) pp255–278. London: Academic Press.

Hannan, E.J. (1969) Identification of vector mixed autoregressive moving average systems. *Biometrika*, **56**, 223–225.

Hannan, E.J. (1970) *Multiple Time Series*. New York: Wiley.

Hannan, E.J. and Diestler, M. (1988) *The Statistical Theory of Linear Systems*. New York: Wiley.

Hanseens, D.M. (1980) Bivariate time series analysis of the relationship between advertising and sales. *Applied Economics*, **12**, 329–339.

Harvey, A.C. (1983) Analysis and generalisation of a multivariate exponential smoothing model. Technical Report, London School of Economics.

Harvey, A.C. (1989) *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

Haugh, L.D. and Box, G.E.P. (1977) Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association*, **72**, 121–130.

Helmer, R.M. and Johansson, J.K. (1977) An exposition of the Box–Jenkins transfer function analysis with an application to the advertising–sales relationship. *Journal of Marketing Research*, **14**, 227–239.

Heuts, R.M.J. and Bronckers, J.H.J.M. (1988) Forecasting the Dutch heavy truck market: a multivariate approach. *International Journal of Forecasting*, **4**, 57–79.

Heyse, J.F. and Wei, W.W.S. (1985a) The partial lag autocorrelation function. Technical Report #32, Department of Statistics, Temple University, PA.

Heyse, J.F. and Wei, W.W.S. (1985b) Modelling the advertising–sales relationship through the use of multiple time series techniques. *Journal of Forecasting*, **4**, 165–181.

Hillmer, S.C. and Tiao, G.C. (1979) Likelihood function of stationary autoregressive moving average models. *Journal of the American Statistical Association*, **74**, 652–660.

Hillmer, S.C. and Tiao, G.C. (1982) An approach to seasonal adjustment. *Journal of the American Statistical Association*, **77**, 63–70.

Hosking, J.R.M. (1980) The multivariate portmanteau statistic. *Journal of the American Statistical Association*, **75**, 602–607.

Jenkins, G.M. (1979) *Practical Experiences with Modelling and Forecasting Time Series*. Jersey: Gwilym Jenkins and Partners Ltd.

Jenkins, G.M. and Alavi, A.S. (1981) Some aspects of modelling and forecasting multivariate time series. *Journal of Time Series Analysis*, **2**, 1–47.

Jenkins, G.M. and McLeod, G. (1982) *Case Studies in Time Series Analysis*, Volume 1. Jersey: Gwilym Jenkins and Partners Ltd.

Johansen, S. (1988) Statistical analysis of co–integration vectors. *Journal of Economic Dynamics and Control*, **12**, 231–254.

Journal of Economic Dynamics and Control (1988) Special issue on Non–stationarity, Co–integration and Error Correction. **12**, (2–3).

Kohn, R. (1982) When is an aggregate of a time series efficiently forecast by its past? *Journal of Econometrics*, **18**, 337–349.

Koreisha, S.G. and Pukkila, T.M. (1987) Identification of non–zero elements in the polynomial matrices of mixed VARMA processes. *Journal of the Royal Statistical Society*, Series B, **41**, 112–126.

Kyle, R.W. (1978) Lydia Pinkham revisited: a Box–Jenkins approach. *Journal of Advertising Research*, **18**, 31–39.

Lin, W.T. (1989) Modelling and forecasting hospital patient movements: univariate and multivariate time series approaches. *International Journal of Forecasting*, **5**, 195–208.

Litterman, R.B. (1986) Forecasting with Bayesian vector autoregressions – five years of experience. *Journal of Business and Economic Statistics*, **4**, 25–38.

Lütkepohl, H. (1984) Linear transformations of vector ARMA processes. *Journal of Econometrics*, **26**, 283–293.

Lütkepohl, H. (1987) *Forecasting Aggregated Vector ARMA Processes*. Berlin: Springer–Verlag.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1984) *The Forecasting Accuracy of Major Time Series Methods*. New York: Wiley.

Mills, T.C. (1990) *Time Series Techniques for Economists*. Cambridge: Cambridge University Press.

Newbold, P. (1988) Some recent developments in time series analysis. III *International Statistical Review*, **56**, 17–29.

Numerical Algorithms Group (1988) Library Manual Mark 13. NAG Ltd, Wilkinson House, Jordan Hill Road, Oxford, OX2 8DR.

Ord, J.K. (1988) Future developments in forecasting. *International Journal of Forecasting*, **4**, 389–401.

Otter, P.W. (1990) Canonical correlation in multivariate time series analysis with an application to one–year–ahead and multiyear–ahead macroeconomic forecasting. *Journal of Business and Economic Statistics*, **8**, 453–458.

Oxford Bulletin of Economics and Statistics (1986) Special issue on Co–integration. **88**, No. 3.

Palda, K.S. (1964) *The Measurement of Cumulative Advertising Effects.* Englewood Cliffs, NJ: Prentice Hall.

Peña, D. and Box, G.E.P. (1987) Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**, 836–843.

Pfeifer, P.E. and Bodily, S.E. (1990) A test of space–time ARMA modelling and forecasting of hotel data. *Journal of Forecasting*, **9**, 255–272.

Phillips, P.C.B. (1987) Time series regression with a unit root. *Econometrica*, **55**, 277–301.

Phillips, P.C.B. and Ouliaris, S. (1988) Testing for co–integration using principal components methods. *Journal of Economic Dynamics and Control*, **12**, 205–230.

Piccolo, D. (1990) A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, **11**, 153–164.

Piccolo, D. and Tunnicliffe Wilson, G. (1984) A unified approach to ARMA model identification and preliminary estimation. *Journal of Time Series Analysis*, **5**, 183–204.

Priestley, M.B. (1981) *Spectral Analysis of Time Series*, Vols. I and II. London: Academic Press.

Priestley, M.B., Subba Rao, T. and Tong, H. (1974) Applications of principal component analysis and factor analysis in the identification of multivariable systems. *IEEE Transactions on Automatic Control*, **AC–19**, 730–734.

Quenouille, M.H. (1968) *The Analysis of Multiple Time Series.* 2nd edn. London: Griffin.

Reilly, D. (1987) *MTS: Multiple time series package.* Automatic Forecasting Systems Inc., PO Box 563, Hatboro, PA 19040.

S–plus (1988) *S–plus for Sun Workstations.* Statistical Sciences U.K. Ltd., 52 Sandfield Road, Headington, Oxford, OX3 7RJ.

Shapiro, D.E. and Switzer, P. (1989) Extracting time trends from multiple monitoring sites. Technical Report #132, Department of Statistics/Department of Applied Earth Sciences, Stanford University, Stanford, CA.

Shea, B. (1987) Estimation of multivariate time series. *Journal of Time Series Analysis*, **8**, 95–109.

Shea, B. (1989) The exact likelihood of a vector autoregressive moving average model. Algorithm AS 242, *Applied Statistics*, **38**, 161–204.

Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.

Solo, V (1986) Topics in advanced time series analysis. in *Lectures in Applied Probability and Statistics*, 1215, 165–328. Berlin: Springer–Verlag.

Spliid, H. (1983) A fast estimation method for the vector autoregressive moving average model with exogenous variables. *Journal of the American Statistical Association*, **78**, 843–849.

Thiel, H. (1966) *Applied Economic Forecasting*. Amsterdam: North–Holland.

Tiao, G.C. and Box, G.E.P. (1981) Modelling multiple time series with applications. *Journal of the American Statistical Association*, **76**, 802–816.

Tiao, G.C. and Guttman, I. (1980) Forecasting contemporal aggregates of multiple time series. *Journal of Econometrics*, **12**, 219–230.

Tiao, G.C. and Tsay, R.S. (1983) Multiple time series modelling and extended sample cross correlations. *Journal of Business and Economic Statistics*, **1**, 43–56.

Tiao, G.C. and Tsay, R.S. (1985) A canonical correlation approach to modelling multivariate time series. *Proceedings of the Economic and Business Section of the American Statistical Association*, 112–120.

Tiao, G.C. and Tsay, R.S. (1989) Model specification in multivariate time series (with Discussion). *Journal of the Royal Statistical Society*, Series B, **51**, 157–213.

Tong, H. (1983) *Threshold Models in Non–Linear Time Series Analysis*. Berlin: Springer–Verlag.

Tsay, R.S. (1986) Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, **81**, 132–141.

Tsay, R.S. (1989a) Parsimonious parameterisation of vector ARMA models. *Journal of Business and Economic Statistics*, **7**, 327–341.

Tsay, R.S. (1989b) Identifying multivariate time series models. *Journal of Time Series Analysis*, **10**, 357–372.

Tunnicliffe Wilson, G. (1973) The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society*, Series B, **35**, 76–85.

Wei, W.W.S. (1990) *Time Series Analysis, Univariate and Multivariate Methods*. Redwood City, CA: Addison–Wesley.

West, M. and Harrison, P.J. (1989) *Bayesian Forecasting and Dynamic Models*. New York: Springer–Verlag.

Wetherill, G.B. (1986) *Regression Analysis with Applications*. London: Chapman and Hall.

Young, P.C. (1984) *Recursive Estimation and Time Series Analysis*. Berlin: Springer–Verlag.

Young, P.C., Jakeman, A. and McMurtries, R. (1980) An instrumental variable method for model order identification. *Automatica*, **16**, 281–294.

Zellner, A. and Palm, F. (1974) Time series analysis and simultaneous equation econometric models. *Journal of Econometrics*, **2**, 17–54.